DISEASE UNIVERSE: VISUALISATION OF POPULATION-WIDE DISEASE-WIDE ASSOCIATIONS

Max Moldovan[1,2]**, Ruslan Enikeev**[3]**, Shabbir Syed-Abdul**[4]**, Phung Anh Nguyen**[4,5]**, Yo-Cheng Chang**[4] **and Yu-Chuan Li**[4]

[1] *Australian Institute of Health Innovation, University of New South Wales, Level 1 AGSM Building, Sydney NSW 2052, Australia;*

[2] *School of Population Health, South Australian Health & Medical Research Institute (SAHMRI), North Terrace, Adelaide, SA 5000;*

[3] *The APAC Sale Group, 3 Petain Rd., #05-07, Singapore 208108;*

[4] *Graduate Institute of Medical Informatics, College of Medical Science and Technology, Taipei Medical University, Taipei, Taiwan;*

[5] *Institute of Biomedical Informatics, National Yang Ming University, Taipei, Taiwan*

Over a lifespan, a human organism is affected by multiple disorders of different origin and severity. We apply a force-directed spring embedding graph layout approach to electronic health records in order to visualise population-wide associations among human disorders as presented in an individual biological organism. The introduced visualisation is implemented on the basis of the Google maps platform and can be found at http://disease-map.net. We argue that the suggested method of visualisation can both validate already known specifics of associations among disorders and identify novel, never noticed association patterns.

Keywords: systems biology; phenomics; electronic health records; visualisation; graph layout

1 INTRODUCTION

It is known that many human disorders are positively associated, accompanying each other due to various, often unknown, genetic, bio-pathological or common risk factors [1]. There is also evidence that some disorders tend to be associated negatively, playing a preventative role against each other, or due to other hypothesised but not properly understood reasons [2, 3]. We use population-wide electronic health records data to visualise how human disorders are positioned against each other in a population with respect to an individual biological organism. By doing so, we attempt to execute a systems biology approach in order to reveal the presence of common functional mechanisms influencing pathogenesis behind groups of human disorders through biological, epidemiological or environmental factors. It is important to note that, due to the specifics of electronic health records [4], together with biological and environmental mechanisms, the method may reflect certain aspects of a healthcare system. For example, closely related but distinct diagnoses are often recorded against the same medical condition. This would induce a positive association between disorders due to healthcare administration rather than biological reasons.

So far the attempts to characterise interactions between human disorders, as observed in a population, have been mainly implemented through network approaches [5, 6]. While being no doubt informative, network approaches predominantly focus on positive association patterns. One of the alternative approaches to the disease-wide association analysis has been reported by Rzhetsky *et al.* [2]. Among other things, the authors objectively characterised probabilities of a person to be affected by a certain disorder, say A, given that the same person has been actually affected by an alternative disorder, say B. This approach allowed us to identify not only positively associated disorders, but also disorders associated negatively – the disorders "competing for the same nucleotide site in the human genome", as hypothesised by the authors. The shortcoming of the study by Rzhetsky *et al.* [2] is that the authors utilised patient records obtained from a single hospital, also covering a limited pre-selected number of diseases. In the following study, we use electronic health records covering the entire population and much wider range of human disorders.

A central objective of the method and its implementation presented below is to visualise association patterns, both positive and negative, among human disorders as observed in an entire population. This would bring to the surface not only already known empirical facts, but also information not previously available. Given that the method implementation can reflect empirical information already known, e.g., a strong positive association between hypertension and diabetes mellitus, as well as unexpected association patterns never noticed before, the presented visualisation can serve as a starting point for formulating novel testable hypotheses in the areas of healthcare and medicine. This can further lead to a better understanding of the complex unobserved dynamics of human disorders intersecting in a single biological body. Such understanding can practically improve the delivery of healthcare and medical treatments.

2 MATERIAL AND METHODS

2.1 Force-directed spring embedding graph layout algorithm

Imagine a single pair of nodes, A and B, positioned on a plane and connected by a spring of a certain *natural* length, δ_{AB}. When the distance between A and B is exactly $d_{AB} = \delta_{AB}$, the spring is in a state of equilibrium, creating neither attraction nor repulsion forces between the nodes (Figure 1a). Moving A and B further apart from each other would create an attraction force (Figure 1b), while moving A and B closer to each other would create a repulsion force between the nodes (Figure 1c).

Given the values of initial required distances δ_{ij} between multiple pairs of nodes, it is rarely possible to locate more than three nodes on a plane such that all required distances between them are satisfied exactly. In fact, it is not even always possible to locate three nodes, keeping the pairwise distances intact, see Figure 2. When distances between the nodes are not satisfied, springs connecting them are not in equilibrium, creating a certain force – either attraction or repulsion.

Aggregated forces created by out-of-equilibrium springs can be expressed by a specific function, drawn from the principle of physics (Hooke's law), leading to system's potential energy U. While a force can be positive (attraction) or negative (repulsion), an energy level is always non-negative irrespective of the sign of the force. Potential energy of a system of M nodes connected by springs of varying stiffness can be expressed as follows:

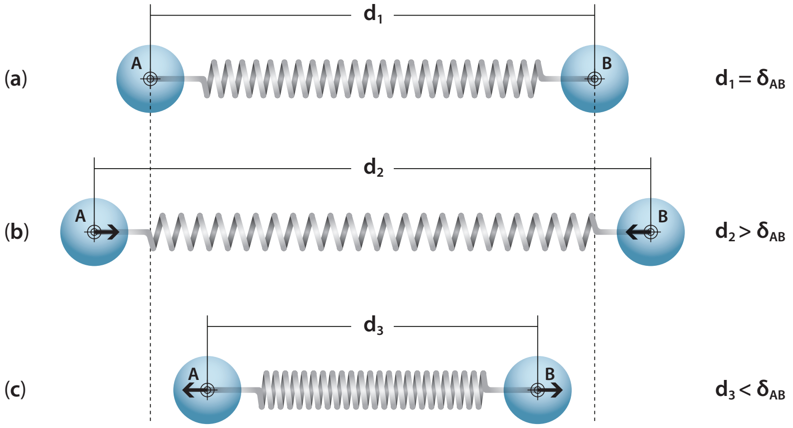


Fig.1 A single pair of nodes in three possible states. **(a)** *Equilibrium*: the nodes neither attract nor repulse; **(b)** *Attraction force*: the nodes are shifted far away from equilibrium and attempt to attract; **(c)** *Repulsion force*: the nodes are closer than if they were in equilibrium and attempt to repulse.

$$U = \frac{1}{2} \sum_{ij \in K} \left((d_{ij} - \widehat{\delta}_{ij})^2 \cdot \kappa_{ij} \right), \qquad K = \binom{M}{2},\ i \neq j \tag{2.1}$$

where $d_{ij} = \sqrt{(X_i - X_j)^2 + (Y_i - Y_j)^2}$ is an Euclidian distance between nodes i and j with coordinates (X_i, Y_i) and (X_j, Y_j), respectively, $\widehat{\delta}_{ij}$ is a natural length of a spring between nodes i and j, $\kappa_{ij} \geq 0$ is an arbitrary parameter that defines the stiffness of a spring between i and j, and K is a number of all possible springs connecting M nodes, with $\binom{\cdot}{\cdot}$ being a binomial coefficient.

By varying pairwise Euclidian distances d_{ij}, the force-directed spring embedding graph layout algorithm [7, 8, 9] performs a search for the configuration of node locations such that system's potential energy U is minimised. By finding the minimum energy U, we attempt to obtain a shape of a system of nodes in which competing forces largely compensate each other. Minimising function (2.1) is a complicated task due to the presence of multiple local minima, and it can rarely be guaranteed that a true global minimum is reached, see Appendix for details. However, we observed that most nodes have nearly constant “designated” locations with respect to other nodes across alternative local minima achieved when minimising (2.1).

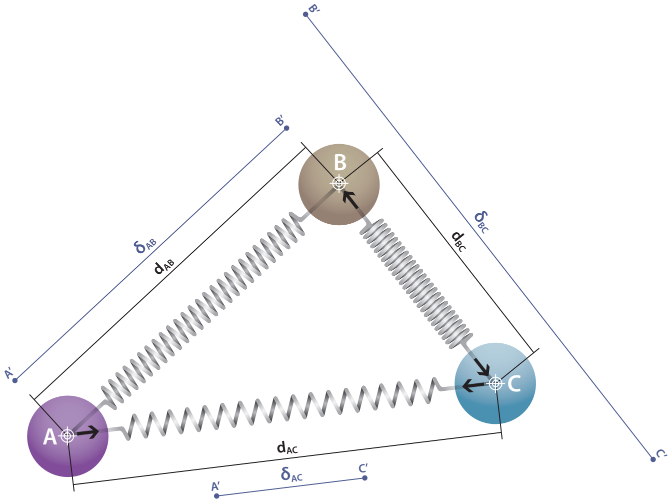


Fig.2 A hypothetical system of three nodes. The initial distances δ_{ij}, $ij \in \{AB, AC, BC\}$ between the nodes are given by the theoretical lines $A'B'$, $B'C'$ and $A'C'$. The joint length of $A'B'$ and $A'C'$ is less than the length of $B'C'$, i.e., $\delta_{AB} + \delta_{AC} < \delta_{BC}$. As a result, for the nodes to connect, one or more of the initial distances between the pairs have to be distorted. The three possible states of springs are equilibrium (AB), attraction (AC) and repulsion (BC).

2.2 Defining natural distances between human disorders

Observing a (sub)-population of size N, suppose that over period T there were C_A individuals with at least one occurrence of disorder A, and C_B individuals with at least one occurrence of disorder B. Further, C_{AB} individuals presented with both disorders A and B, each disorder observed at least once over the same period. Then the information can be summarised as shown by Table 1.

Table 1: Occurrence counts of disorders A and B in population of size N.

	Disorder A		
Disorder B	A present	A absent	Total
B present	C_{AB}	$\cdot$	C_B
B absent	$\cdot$	$\cdot$	–
Total	C_A	–	N

Table 1 is an example of a 2×2 table with fixed margins. Assuming that individuals are affected independently of each other (which can be violated, e.g., for infectious diseases), X follows a non-central hypergeometric distribution $X \sim \text{Hyper}(N, C_A, C_B)$ given by

[10, 11]:

$$\Pr(X = C_{AB}) = \frac{\binom{C_B}{C_{AB}}\binom{N-C_B}{C_A-C_{AB}}}{\binom{N}{C_A}} e^{\theta_{AB} C_{AB}} \tag{2.2}$$

where $\max(0, C_A + C_B - N) \leq \mathbf{C_{AB}} \leq \min(C_A, C_B)$, $\theta \in (-\infty, +\infty)$ is a log-odds ratio, and $e = 2.718\ldots$ is the base of a natural logarithm. For algorithm implementation, conditional maximum likelihood estimates of θ were approximated by unconditional log-odds ratios:

$$\widehat{\theta}_{AB} = \ln\left(\frac{C_{AB}(N - C_A + C_{AB} - C_B)}{(C_A - C_{AB})(-C_{AB} + C_B)}\right) \tag{2.3}$$

where $\ln(\cdot)$ is a natural logarithm. Switching the risk factor from being B for A to being A for B does not effect log-odds estimates. Natural (equilibrium) lengths of springs between nodes i and j were obtained through the following *reversed* expit transform [10, p.121]:

$$\widehat{\delta}_{ij} = \frac{\exp(-\widehat{\theta}_{ij})}{1 + \exp(-\widehat{\theta}_{ij})} \tag{2.4}$$

where $\widehat{\delta}_{ij} \in [0, 1]$ by construction. Note that the sign on log-odds estimate $\widehat{\theta}$ was changed to the opposite (i.e., reversed), making stronger positive associations correspond to the smaller values of $\widehat{\delta}_{ij}$. We do so in order to make $\widehat{\delta}_{ij}$ resemble Euclidian distances between the nodes.

Due to estimation, there is uncertainty about $\widehat{\delta}$ values obtained from the data. Such an uncertainty is usually handled by reporting confidence intervals corresponding to $\widehat{\delta}$. In the present version of method implementation, we intentionally avoided using confidence intervals, p-values or other statistical tools normally involved in hypothesis testing. We did so in order to reflect the empirical information contained in the data without any subjective interpretation that could otherwise be introduced through, for example, the choice of a significance level.

2.3 Potential alternative implementations

It should be noted that the force-directed spring embedding graph layout method for visualising relationships among multiple objects, human disorders in our case, is not the only approach available. Together with network algorithms already mentioned above, multidimensional scaling [12] and biplot [13] methods are two more approaches for two-dimensional visualisation of relationships between multiple objects. However, it is important to emphasise that both methods, at least in their traditional form, focus on similarities between objects, which would correspond to positive associations between disorders as per our current settings. While an application of alternative methods to our empirical data set would be interesting and potentially informative, multidimensional scaling and biplot methods are unlikely to address negative associations between disorders in a proper way.

At the same time, Euclidian distances between disorders, as specified by (2.4) in our study, could be easily reversed, i.e., positive and negative associations can be made corresponding to longer and shorter Euclidian distances between the nodes in Figure 1, respectively. Such an alternative vision of the problem could reveal an entirely new set of empirical information, not reflected by the current method implementation. We retain this research direction for later investigation.

3 PRACTICAL IMPLEMENTATION

3.1 Empirical information

The presented visualisation has been motivated by the Internet Map implementation [14]. We use electronic health records obtained from the Taiwanese national health insurance research database covering the entire population of Taiwan over the period of three years (2000-2002). The same three-year observation window of the maximum available length has been used to record the counts corresponding to Table 1. Disorder records are based on ICD9-CM (International Classification of Diseases, Ninth Revision, Clinical Modification), five-digit version. The dataset has been stratified into male and female groups. Each of these two groups has been further stratified into ten age sub-groups, i.e., 0-9, 10-19, . . . , 90+. Each subject within a sub-group was noted by his of her first insurance claim starting from 01 January 2000, assigned to a certain age-gender group and followed for the rest of the period ending on 31 December 2002. Codes corresponding to E and V categories of ICD9-CM (External causes of injury and Supplemental classification) were excluded from consideration.

3.2 Prevalence threshold and spring stiffness

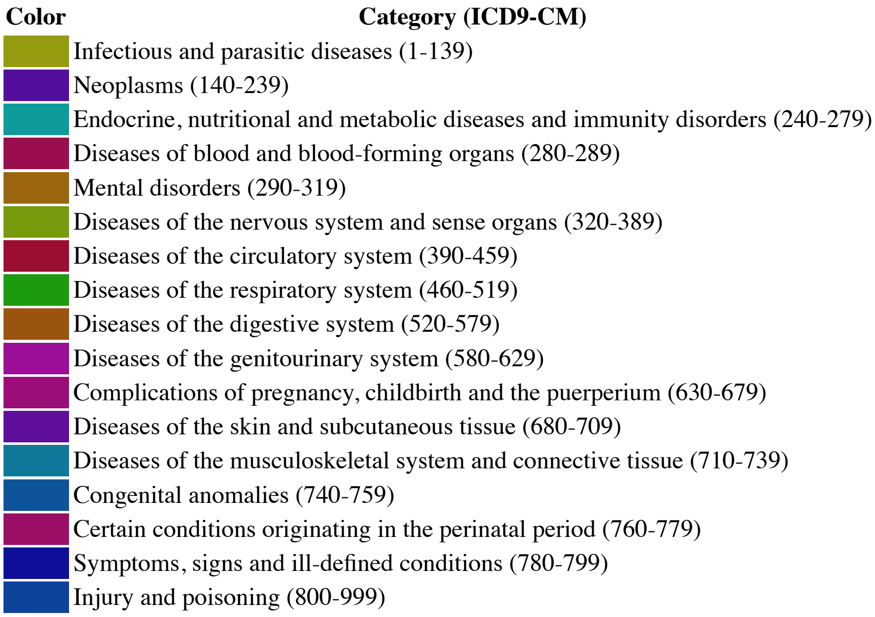
We compute $\widehat{\delta}_{ij}$ given by (2.4) for each observed pair of disorders i and j. An empirical examination of $\widehat{\delta}_{ij}$ revealed that log-odds estimates $\widehat{\theta}_{ij}$ that underlie $\widehat{\delta}_{ij}$, exhibit anomalous behaviour for smaller counts C_i and C_j, i.e., they tend to be much larger than it would be expected under a random process. Such an anomaly would bias the attention of an optimisation algorithm applied to (2.1) towards diseases of a smaller prevalence. We attributed this anomaly to exceptionally high positive associations between certain low prevalence pairs of disorders as observed in the context of the entire population and reflected by odds ratio estimates. In particular, the expected value of X in the hypergeometric distribution function (2.2) when $\theta = 0$, i.e., there is no association between disorders, is given by [11, p.93]:

$$REC_{ij} = \frac{C_i C_j}{N} \tag{3.5}$$

where REC stands for Random Expected Co-occurrence. We interpret REC_{ij} as a value reflecting "visibility" of co-occurrences between i and j, with higher visibility (i.e., greater values of REC_{ij}) leading to more reliable empirical outcomes C_{ij} in Table 1. Keeping this interpretation in mind, we have executed the following ad hoc solution for dealing with the identified anomaly. Firstly, we imposed the threshold $C = \sqrt{2N}$ on disease occurrence counts. This guarantees that $REC_{ij} > 2$ for all possible C_i and C_j. The meaning behind this restriction is to ensure that only theoretically "visible" $\widehat{\theta}_{ij}$ estimates are used for visualisation. The cost is that we dismissed smaller prevalence disorders that never exceeded $REC_{ij} = 2$ in any of the age-gender groups. Secondly to the imposed lower limit on the observed occurrence counts, we set the stiffness parameter of a spring between pairs i and j to $\kappa_{ij} = \ln(REC_{ij})$. This modification makes sure that less theoretically “visible” co-occurrences C_{ij} are given less importance when minimising the energy function (2.1).

3.3 Visualisation

The Google maps platform (https://developers.google.com/maps/) was used to visualise the outcomes. The sizes of the nodes are set to be proportional to the observed disease prevalence in the corresponding age-gender stratified sub-groups. The colour codes of the nodes correspond to the broad disease categories as per ICD9-CM classification, see Figure 3. All maps are displayed in the same coordinate system with the same scale so that they can be compared against each other.



Color	Category (ICD9-CM)
🟨	Infectious and parasitic diseases (1-139)
🟪	Neoplasms (140-239)
🟦	Endocrine, nutritional and metabolic diseases and immunity disorders (240-279)
🟥	Diseases of blood and blood-forming organs (280-289)
🟫	Mental disorders (290-319)
🟩	Diseases of the nervous system and sense organs (320-389)
🟥	Diseases of the circulatory system (390-459)
🟩	Diseases of the respiratory system (460-519)
🟫	Diseases of the digestive system (520-579)
🟪	Diseases of the genitourinary system (580-629)
🟪	Complications of pregnancy, childbirth and the puerperium (630-679)
🟪	Diseases of the skin and subcutaneous tissue (680-709)
🟦	Diseases of the musculoskeletal system and connective tissue (710-739)
🟦	Congenital anomalies (740-759)
🟪	Certain conditions originating in the perinatal period (760-779)
🟦	Symptoms, signs and ill-defined conditions (780-799)
🟦	Injury and poisoning (800-999)

Fig.3 Broad disease categories as per ICD9-CM classification and the corresponding colour codes as displayed at http://disease-map.net.

4 SOME EXAMPLES OF USING THE MAPS

4.1 An accidental proximity?

When exploring the maps, some regions and mutual locations can attract attention due to certain, often subjective, reasons. For example, observing the map for females age 40-49 (F40+), in the central region towards south-east we find that *peptic ulcer* (ICD9-CM 533) is located side by side with *neurotic disorders* (ICD9-CM 300). Have these disorders fallen close together by chance? A search through the medical literature has quickly identified that an abnormal association between peptic ulcer and neurotic disorders was noticed years ago [15, 16]. Looking at a wider category of digestive system disorders (ICD9-CM 520 to 579), this mutual location pattern remains largely the same over multiple maps (e.g., see M30+, M40+ and F30+), leading to several testable hypotheses. One example of such a hypothesis would be: “There is no direct psychosomatic link between digestive system disorders and neurotic disorders”.

Viewed from a different angle, the literature coverage on pairs of closely located disorders appears not to be accidental. Syed-Abdul *et al.* [17] found that the proximity of disorder pairs is positively correlated with the degree of literature coverage, the latter being represented by a number of hits for a pair of disorders returned from an appropriate query to the PubMed search engine.

4.2 Unlikely neighbours: a cancer-schizophrenia association puzzle

The topic of observed evidence of associations between schizophrenia and various cancers has been widely debated [18]. Evidence tends to point towards the presence of a negative association between schizophrenia and several cancers, even though there is no absolute consensus [3, 19]. A shared genetic architecture has been proposed as a reason for observed associations [20, 21]. Alternatively, there is evidence that the chances of schizophrenia patients being timely diagnosed with certain types of cancer, on average, are lower than for general population non-schizophrenic patients [22].

Exploring the maps, it can be found that *schizophrenic disorders* (ICD9-CM 295) consistently fall on the southern border of the maps, sometimes being the most distant points from the imaginary centre of a “galaxy”, e.g., see M50+. Interestingly, various types of cancers also regularly fall on the same southern border even though, consistently with the literature, association estimates for schizophrenia-cancer pairs regularly cross to the negative side, i.e., $\widehat{\delta}_{ij}$ given by (2.4) exceeds the value of 0.5.

One potential explanation for such an anomaly would be that schizophrenia and some cancers have closely related underlying causes revealed through similar relationships with other disorders. In this way, schizophrenia and cancers are placed to their southern border locations by the forces generated within the system. Often being negatively associated, schizophrenia and cancers are like two sides of the same coin, “competing for the same nucleotide site in the human genome” as per Rzhetsky *et al.* [2] vision, but potentially disassociated due to deeper, not properly recognised and understood reasons which are still to be identified and investigated. The visualisation we introduced is a tool for originating and directing such investigations.

5 CONCLUSION

Electronic health records have become an integral part of national healthcare systems worldwide, and it is essential to comprehensively utilise the information contained in the growing number of databases. The method we introduced is one of the effective and informative tools for doing so. While the current realisation of the method has its obvious limitations, the presented maps are the first implementation of this kind and intended to set a reference benchmark for further developments in the same direction. A formal empirical validation of the introduced visualisation is beyond the scope of this paper, but based on the broad examination of the resulted maps, we argue that the presented implementation can both assist with validation of already known phenomena as well as with identification of novel, previously never noticed, association patterns related to functional aspects of medicine and healthcare. We suggest that the maps be used for generating testable hypotheses and invite the reader to explore the vast amount of information contained in them at http://disease-map.net.

ACKNOWLEDGMENTS: We are grateful to Chris Lloyd and Nicholas Nechval who provided their constructive feedback that helped to improve the manuscript. We thank Hanna Kalkova and Andrey Stepanov for preparing Figures 1 and 2.

AUTHOR CONTRIBUTIONS: M.M., R.E., S.S-A. and Y-C.L. invented and developed the concept (IPN-13-000062, NewSouth Innovations). S.S-A., M.M., Y-C.L. and Y-C.C. performed selective biological validation. R.E. wrote software, implemented the method and developed the website. M.M. and R.E. wrote the manuscript. Y-C.L. obtained the data. PA.N. organised and validated the data.

FUNDING: No formal targeted funding was allocated towards the project.

COMPETING FINANCIAL INTERESTS: The authors declared no competing financial interests.

References

[1] Ferrannini, E. and Cushman, W.C. (2012) “Diabetes and hypertension: the bad companions.” *Lancet*, 380, 601-610.

[2] Rzhetsky, A, Wajngurt, D., Park, N., and Zheng, T. (2007) “Probing genetic overlap among human phenotypes.” *Proceedings of the National Academy of Sciences*, 104, 11694-11699.

[3] Chou, F.H.-C., Tsai, K.-Y., Su, C.-Y., and Lee, C.-C. (2011) “The incidence and relative risk factors for developing cancer among patients with schizophrenia: A nine-year follow-up study.” *Schizophrenia Research*, 129, 97-103.

[4] Hripcsak, G. and Albers, D.J. (2013) “Next-generation phenotyping of electronic health records.” *Journal of the American Medical Informatics Association*, 20, 117-121.

[5] Hidalgo, C.A., Blumm. N., Barabási, A.-L., and Christakis, N.A. (2009) “A dynamic network approach for the study of human phenotypes.” *PLoS Computational Biology*, 5(4), e1000353.

[6] Barabási, A.-L., Gulbahce, N., and Loscalzo, J. (2011) “Network medicine: a network-based approach to human disease.” *Nature Reviews - Genetics*, 12, 56-68.

[7] Kamada, T. and Kawai, S. (1989) “An algorithm for drawing general undirected graphs.” *Information Processing Letters*, 31, 7-15.

[8] Tunkelang, D. (1999) “A numerical optimisation approach to general graph drawing.” PhD thesis, Carnegie Mellon University: http://reports-archive.adm.cs.cmu.edu/anon/1998/CMU-CS-98-189.pdf

[9] Hu, Y. (2005) “Efficient, high-quality force-directed graph drawing.” *The Mathematica Journal*, 10, 37-71.

[10] Lloyd, C.J. (1999) *Statistical Analysis of Categorical Data*. Wiley (New York).

[11] Agresti, A. (2002) *Categorical Data Analysis*. 2d edition, John Wiley & Sons.

[12] Cox, T.F. and Cox, M.A.A. (2001) *Multidimensional Scaling.* Chapman and Hall.

[13] Gabriel, K.R. (1971) “The biplot graphic display of matrices with application to principal component analysis.” *Biometrika*, 58(3), 453-467.

[14] Enikeev, R. (2012) *The Internet map.* Singapore, http://internet-map.net.

[15] Montgomery, H., Schindler, R., Underdahl, L.O., Butt, H.R., and Walters, W. (1944) “Peptic ulcer, gastritis and psychoneurosis.” *Journal of the American Medical Association*, 125, 890-894.

[16] Stafford-Clark, D. (1952) “Peptic ulcer and the neurotic.” *British Medical Journal*, 16, 390-391.

[17] Syed-Abdul, S., Enikeev, R., Moldovan, M., Jian, W.-S., Nguyen, A., Iqbal, U., Chang, Y.-C., Hsu, M.-H., Lin, S.C., and Li, Y.-C. (2013) Capturing and visualizing human diseasomic associations: A population-based observational study. Manuscript.

[18] Hodgson, R., Wildgust, H.J., and Bushe, C.J. (2010) Cancer and schizophrenia: is there a paradox? *Journal of Psychopharmacology*, 24, 51-60.

[19] Hippisley-Cox, J., Vinogradova, Y., Coupland, C., and Parker, C. (2007) “Risk of malignancy in patients with schizophrenia or bipolar disorder.” *Archives of General Psychiatry*, 64, 1368-1376.

[20] Catts V.S., Catts S.V., O‘Toole B.I., and Frost, A.D.J. (2008) “Cancer incidence in patients with schizophrenia and their first-degree relatives – a meta-analysis.” *Acta Psychiatrica Scandinavica*, 117, 323-336.

[21] Gal, G., Goral, A., Murad, H., Gross, R., Pugachova, I., Barchana, M., Kohn, R., and Levav, I. (2012) “Cancer in parents of persons with schizophrenia: Is there a genetic protection?” *Schizophrenia Research*, 139, 189-193.

[22] Crump, C., Winkleby, M.A., Sundquist, K., and Sundquist, J. (2013) “Comorbidities and mortality in persons with schizophrenia: a Swedish national cohort study.” *American Journal of Psychiatry*, 170, 324-333.

Corresponding authors
Max Moldovan (max.moldovan@gmail.com) and Yu-Chuan Li (jaak88@gmail.com)

Appendix: Energy minimisation method

Finding a global minimum of (2.1) is a complicated task due to the presence of multiple local minima of this function. Different approaches of global minimisation can be applied, but it can be rarely known when and if the global minimum is reached, unless a minimum energy level is known in advance. Our current implementation of energy minimisation is to use multiple local searches with the conjugate gradient algorithm from random starting positions in order to obtain a *master* map – the map that includes diseases across the entire spectrum of age groups and both genders. In each of multiple attempts, the nodes are dropped on the map with random positions (X, Y), and the conjugate gradient algorithm runs searching for the closest local minimum of U in (2.1). If the new local minimum is less than the best (smallest) minimum recorded over previous attempts, it becomes the new best minimum. The procedure is repeated until the best minimum stops changing even after a reasonably large (4000, in our implementation) number of random allocation attempts, see Algorithm 1. The computational complexity of the algorithm is $O(n^2)$.

The obtained master map served as a collection of starting points for the age-gender stratified maps, see Algorithm 2. Minimising (2.1) from a single set of starting points leads to a local minimum that can be further improved by applying the minimisation approach used for obtaining the master map. However, we still used minimisation from the single set of starting points in order to make the maps comparable across age groups and genders. Table A1 reports the achieved minimum energy levels using "partial" minimisation as per Algorithm 2 compared to the "full" minimisation implemented through Algorithm 1.

Table A1 Minimum achieved energy levels from partial and (attempted) full minimisation approaches.

Group	Subjects followed (N)	Disorder numbers	Partial	Full	Per cent improve
F 0-9	1,677,365	565	7,807.96	7,700.69	1.39
F 10-19	1,595,057	743	9,470.09	9,166.34	3.31
F 20-29	1,780,095	1041	22,897.04	22,268.39	2.82
F 30-39	1,765,866	1136	25,914.10	25,387.60	2.07
F 40-49	1,631,968	1243	31,126.22	30,913.37	0.69
F 50-59	930,496	1251	33,451.35	33,334.80	0.35
F 60-69	711,096	1271	36,129.76	36,056.87	0.20
F 70-79	427,821	1177	29,935.15	29,857.36	0.26
F 80-89	141,225	783	10,802.72	10,773.33	0.27
F 90-99	8,532	176	318.26	315.74	0.80
M 0-9	1,827,447	630	10,068.44	9,910.12	1.60
M 10-19	1,678,415	721	9,451.03	9,346.16	1.12
M 20-29	1,767,163	859	12,532.53	12,345.32	1.52
M 30-39	1,737,715	948	14,263.07	14,099.18	1.16
M 40-49	1,577,320	1090	19,485.22	19,454.02	0.16
M 50-59	898,150	1065	20,296.22	20,247.20	0.24
M 60-69	692,061	1163	26,737.58	26,563.97	0.65
M 70-79	532,308	1225	30,740.90	30,622.10	0.39
M 80-89	133,480	781	10,636.67	10,599.66	0.35
M 90-99	4,769	151	240.25	238.70	0.65
Master	21,518,574	2298	–	130,381.91	–

Algorithm 1 Energy minimisation for the master map.

```
Require: $\gamma \leftarrow 0.01$    /* tolerance for the change in objective function (2.1)
Require: $s \leftarrow 1$    /* initial step size
Require: $\tau \leftarrow 0.9$    /* step decrease rate
Require: $s_{min} \leftarrow 0.000001$    /* minimum step tolerance
Require: $\widehat{\delta}_{ij}$ for $K = \binom{M}{2}$ pairs, $i \neq j$    /* pairwise natural lengths given by (2.4)
Require: $U_{current} \leftarrow +Inf$    /* current energy level to be reduced
Require: $cc \leftarrow 0$    /* random positions attempts counter
Require: $cc_{max} \leftarrow 4000$    /* maximum number of attempts with no energy reduction
  while $(cc < cc_{max})$ do
    $(X_0, Y_0) \leftarrow random()$    /* drop nodes at random positions
    $U_0 \leftarrow f_E(X_0, Y_0)$    /* value of objective function (2.1)
    $G \leftarrow \{-\nabla\,(f_E(X_0, Y_0))\}$    /* define antigradients for the first step
    $(\Delta X, \Delta Y) \leftarrow f_G(G)$    /* step direction
    $(X, Y) \leftarrow (X_0, Y_0) + (\Delta X, \Delta Y) \cdot s$    /* current coordinates of nodes
    $U \leftarrow f_E(X, Y)$    /* current value of objective function (2.1)
    $\Delta U \leftarrow (U_0 - U)$    /* change in energy
    while $(\Delta U > \gamma)$ & $(s > s_{min})$ do
      $G_C \leftarrow \{\nabla_C\,(f_E(X_0, Y_0; X, Y))\}$    /* evaluate conjugate gradients
      $(\Delta X, \Delta Y) \leftarrow f_{CG}(G_C)$    /* step direction
      $(X_{temp}, Y_{temp}) \leftarrow (X, Y) + (\Delta X, \Delta Y) \cdot s$    /* trial coordinates of nodes
      $U \leftarrow f_E(X_{temp}, Y_{temp})$    /* current value of the objective function
      if $U < U_0$ then
        $\Delta U \leftarrow (U_0 - U)$    /* update change in energy
        $U_0 \leftarrow U$    /* update preceding energy value
        $(X_0, Y_0) \leftarrow (X, Y)$    /* update preceding coordinates
        $(X, Y) \leftarrow (X_{temp}, Y_{temp})$    /* assign the values of current coordinates
      else
        $s \leftarrow s \cdot \tau$    /* reduce step size
      end if
    end while
    if $U_0 < U_{current}$ then
      $U_{current} \leftarrow U_0$    /* update minimum energy value
      $(X_{current}, Y_{current}) \leftarrow (X, Y)$    /* update coordinates
      $cc \leftarrow 0$    /* set attempts count to zero
    else
      $cc \leftarrow cc + 1$    /* next attempt
    end if
  end while
  $(X_{master}, Y_{master}) \leftarrow (X_{current}, Y_{current})$
  return $(X_{master}, Y_{master})$    /* nodes' coordinates under minimum energy achieved
```

Algorithm 2 Energy minimisation for age and gender stratified maps.

```
Require: $\gamma \leftarrow 1e\text{-}5$    /* tolerance for the change in objective function (2.1)
Require: $s \leftarrow 1$    /* initial step size
Require: $\tau \leftarrow 0.9$    /* step decrease rate
Require: $s_{min} \leftarrow 0.000001$    /* minimum step tolerance
Require: $\widehat{\delta}_{ij}$ for $K = \binom{M}{2}$ pairs, $i \neq j$    /* pairwise natural lengths given by (2.4)
  $(X_0, Y_0) \leftarrow (X_{master}, Y_{master})$    /* use coordinates from the master map as starting points
  $U_0 \leftarrow f_E(X_0, Y_0)$    /* the value of objective function (2.1)
  $G \leftarrow \{-\nabla\,(f_E(X_0, Y_0))\}$    /* define antigradients for the first step
  $(\Delta X, \Delta Y) \leftarrow f_G(G)$    /* step direction
  $(X, Y) \leftarrow (X_0, Y_0) + (\Delta X, \Delta Y) \cdot s$    /* current coordinates of nodes
  $U \leftarrow f_E(X, Y)$    /* current value of objective function (2.1)
  $\Delta U \leftarrow (U_0 - U)$    /* change in energy
  while $(\Delta U > \gamma)$ & $(s > s_{min})$ do
    $G_C \leftarrow \{\nabla_C\,(f_E(X_0, Y_0; X, Y))\}$    /* evaluate conjugate gradients
    $(\Delta X, \Delta Y) \leftarrow f_{CG}(G_C)$    /* step direction
    $(X_{temp}, Y_{temp}) \leftarrow (X, Y) + (\Delta X, \Delta Y) \cdot s$    /* trial coordinates of nodes
    $U \leftarrow f_E(X_{temp}, Y_{temp})$    /* current value of the objective function
    if $U < U_0$ then
      $\Delta U \leftarrow (U_0 - U)$    /* update change in energy
      $U_0 \leftarrow U$    /* update preceding energy value
      $(X_0, Y_0) \leftarrow (X, Y)$    /* update preceding coordinates
      $(X, Y) \leftarrow (X_{temp}, Y_{temp})$    /* assign values of current coordinates
    else
      $s \leftarrow s \cdot \tau$    /* reduce step size
    end if
  end while
  $(X_{stratif}, Y_{stratif}) \leftarrow (X, Y)$
  return $(X_{stratif}, Y_{stratif})$    /* nodes' coordinates under minimum energy achieved
```