Reviewers' Comments:

Reviewer #1:
Remarks to the Author:
Review for:

"A Scalable, Web-based, and Quantitative Social Network Assessment for Identifying Potentially Modifiable Risks in the Social Environment"
Amar Dhand, Charles C. White, Catherine Johnson, Zongqi Xia, Philip De Jager

In this paper, the authors present a new quantitative social network assessment tool on a secure open source web platform that could be helpful for large scale clinical studies. They use the tool for demonstration to quantify the social network structure and composition of 1493 persons in US with a risk to develop multiple sclerosis. While network structure measures were not associated with self-reported neurological disabilities, negative health habits of individual's surroundings strongly correlated with the neurological disorder.
There are several limitations of the current study including lack of causal conclusions, self-reporting of both network characteristics and neurological functions and in terms of generalization of the results. On the other hand, the paper is well written, the study technically sounds and the topic is interesting in the broad area of designing interventions and evaluating clinical studies. For those reasons, even if I cannot be very enthusiastic, I believe an improved version could be accepted for publication in Nature Communications. I have some remarks that need to be addressed:

1. Figure 1 (the last part in the right) is not very informative. While the reader understands from the text the separation between the structure and composition, Figure 1 at the right get things complicated.
2. In lines 94-96 the authors state that "the maximum and mean degree … are indicating the distribution of ties in the network." Usually social networks are complex in the sense that the degree distribution follows a heterogeneous distribution. Therefore, maximum and mean degree is not conclusive for tie distribution in a network. I propose that the authors change a bit the wording here.
3. Figure 2 is not informative and add nothing to the contribution of the study. It is a very simple scatter plot of the locations of individual participants that gets very complicated in large cities (circles are one over the other) – and it is difficult to fit in a Nature Communication paper. I would have this figure in the supplementary material rather than in the main text. It looks like that the location information is self-reported through an address. Maybe a heat map at the level of county or zip code would look much better. It looks like there is a large correlation between population of a city (or county) and the number of participants. This is something that the authors could report.
4. In line 142 the authors refer to Fig 3. It should be Fig 4. Also in the legend of Fig 4 it refers to "orange" color. I believe the unhealthy friends are red colored.
5. In Figure 3, it looks like that about ~25% of all ego's networks have healthy habits. I think is worth mentioning the exact percentage. Also, is the size of the network correlated (at least weakly) with how "unhealthy" is the network? It should be that the more the friends I have the largest the probability that at least one of them has unhealthy habits. Therefore, the way networks are ranked in Figure 3 is related to the way that networks in Figure 4 are ranked and visualized.
6. My above comment on the correlation between network size and percentage of friends that follow unhealthy habits is the main reason (from my point of view) behind the result that the authors report in Table 3, i.e. the association between the MSRS and the size of the network. There is this confounding here that the structure (at least the size) is highly correlated with the network composition - health. This is something that needs to be discussed in the paper – maybe in page 11.
7. At lines 232-247 the authors discuss about "homophily". It is a common sense that association (correlation) of health behaviors in a social network can be attributed to homophily (the tendency of similar individuals to connect in a social network) or to contagion effects, or to the fact that

certain areas of a network subject to same externalities. In line 234 the authors state that a "potential mechanism of homophily is social contagion." I would say that potential mechanism of the association in health behaviors are homophily and social contagion.

Reviewer #2:
Remarks to the Author:
This is a very well written and interesting paper that looks at a novel metric that may impact on health. As such it provides new information that may be used to influence the risk of developing MS in those at risk. The social network is certainly a very topical area for research and the authors address this in the GEMS cohort.
I have several questions.
Firstly the MSRS_R was used as the primary outome measure of disability but was heavily skewed to a zero resposne. Di the authors consider using a hurdle model to anlyse the data??
What was the influence of educational level on social networks could this be a confounder?
In table 2 it would be useful to provide the actual percentages who smoke etc in addition to the medians IQR.
Could the authors also comment on how representative their sample was of the whole GEMS cohort would it be more likely that those who respond have a greater or lesser social network than those who did not?
The information presented here would defintiley allow me to reproduce this work.
Bruce taylor
University of Tasmania

Response to the Referees' Comments

In this paper, the authors present a new quantitative social network assessment tool on a secure open source web platform that could be helpful for large scale clinical studies. They use the tool for demonstration to quantify the social network structure and composition of 1493 persons in US with a risk to develop multiple sclerosis. While network structure measures were not associated with self-reported neurological disabilities, negative health habits of individual's surroundings strongly correlated with the neurological disorder.

There are several limitations of the current study including lack of causal conclusions, self-reporting of both network characteristics and neurological functions and in terms of generalization of the results. On the other hand, the paper is well written, the study technically sounds and the topic is interesting in the broad area of designing interventions and evaluating clinical studies. For those reasons, even if I cannot be very enthusiastic, I believe an improved version could be accepted for publication in *Nature Communications*. I have some remarks that need to be addressed:

1. Figure 1 (the last part in the right) is not very informative. While the reader understands from the text the separation between the structure and composition, Figure 1 at the right get things complicated.

We have simplified the right side of Figure 1. We have removed two graphs and focused the reader on two modifications possible: adding a new friend or improving the habits of unhealthy social contacts.

2. In lines 94-96 the authors state that "the maximum and mean degree … are indicating the distribution of ties in the network." Usually social networks are complex in the sense that the degree distribution follows a heterogeneous distribution. Therefore, maximum and mean degree is not conclusive for tie distribution in a network. I propose that the authors change a bit the wording here.

We have removed this wording and changed the sentence as follows:

P.5, line 99: Maximum and mean degree are the network members who have the most number of ties and average number of ties, respectively.

3. Figure 2 is not informative and add nothing to the contribution of the study. It is a very simple scatter plot of the locations of individual participants that gets very complicated in large cities (circles are one over the other) – and it is difficult to fit in a Nature Communication paper. I would have this figure in the supplementary material rather than in the main text. It looks like that the location information is self-reported through an address. Maybe a heat map at the level of county or zip code would look much better. It looks like there is a large correlation between population of a city (or county) and the number of participants. This is something that the authors could report.

We have moved the figure to supplementary material. The location of participants is interesting but not directly related to the point of the paper, so we do not add these details into the manuscript. We will consider this variable for future analyses.

4. In line 142 the authors refer to Fig 3. It should be Fig 4. Also in the legend of Fig 4 it refers to "orange" color. I believe the unhealthy friends are red colored.

The figure numbers are corrected. The color description is corrected in the legend.

P.28, line 742: Fig 3: A montage of the social network composition of all participants with respect to healthy habits around the participant. Red dots are persons in the network with a negative health influence…

5. In Figure 3, it looks like that about ~25% of all ego's networks have healthy habits. I think is worth mentioning the exact percentage. Also, is the size of the network correlated (at least weakly) with how "unhealthy" is the network? It should be that the more the friends I have the largest the probability that at least one of them has unhealthy habits. Therefore, the way networks are ranked in Figure 3 is related to the way that networks in Figure 4 are ranked and visualized.

We found that 17% of participants had networks in which all persons were healthy. We added this result in the text. There is a weak negative correlation between network size and proportion of unhealthy network members (Pearson's correlation = -0.13, 95% CI (-0.08 – -0.17)). We added this result to the text.

P.7, line 153: 17% of participants had personal networks in which all members were healthy.…There was a weak negative correlation between network size and the percentage of network members with unhealthy habits (Pearson's correlation = -0.13, 95% confidence interval (-0.08 – -0.17)).

6. My above comment on the correlation between network size and percentage of friends that follow unhealthy habits is the main reason (from my point of view) behind the result that the authors report in Table 3, i.e. the association between the MSRS and the size of the network. There is this confounding here that the structure (at least the size) is highly correlated with the network composition - health. This is something that needs to be discussed in the paper – maybe in page 11.

We had the same concern, and addressed it by dividing all compositional trait frequencies by network size to create the percentage variables. Therefore, the percentage of persons who don't exercise are the number of network members who don't exercise divided by network size. Therefore, we believe our analyses account for size. We emphasize this on page 5.

P.5, line 110: All compositional variables were created to account for network size. Specifically, the number who fit a category were divided by the total size to create the percentage.

7.   At lines 232-247 the authors discuss about "homophily". It is a common sense that association (correlation) of health behaviors in a social network can be attributed to homophily (the tendency of similar individuals to connect in a social network) or to contagion effects, or to the fact that certain areas of a network subject to same externalities. In line 234 the authors state that a "potential mechanism of homophily is social contagion." I would say that potential mechanism of the association in health behaviors are homophily and social contagion.

We changed this wording throughout the paragraph to not overlap with the prior paragraph that already discusses homophily and instead focus on social contagion and externalities.

P.11, line 256: Two more mechanisms that may explain the association of network members' health habits and the participant's neurological disability are social contagion and antecedent exposures. Social contagion is a type of social influence in which behavior in one or many network members affects the behavior of the index participant. Detection of this effect requires longitudinal data and network modeling, such as stochastic actor-oriented or instrumental variable approaches, to understand the spread of behaviors through social ties. For example, one study shows the spread of physical activity in 1 million users of a smartphone running app[31]. Antecedent exposures influencing both parties may be another contributor. For example, rural environments with poor access to medical services may influence the habits of all members of the network with regard to seeking medical care. Finally, a combination of these factors may explain the association of poor health habits in the network and a person's neurological disability.

Reviewer 2's comments:

This is a very well written and interesting paper that looks at a novel metric that may impact on health. As such it provides new information that may be used to influence the risk of developing MS in those at risk. The social network is certainly a very topical area for research and the authors address this in the GEMS cohort.
I have several questions.

1. Firstly the MSRS_R was used as the primary outcome measure of disability but was heavily skewed to a zero response. Did the authors consider using a hurdle model to analyze the data?

We completed a hurdle model with both poison regression and negative binomial regression to analyze the data. We found they were similar to our non-parametric spearman rank associations (see Supplementary Table 1). The spearman rank associations are robust to non-normal distributions and are easy to interpret. Therefore, after considering all the models, we kept the non-parametric spearman rank associations as the main sensitivity analysis.

2. What was the influence of educational level on social networks could this be a confounder?

We agree that years of education could be a confounder. Therefore, we completed additional analyses and revised our linear regression analyses now adjusting for age, sex, marital status, and years of education. The two strongest network composition variables, percent of network members (1) who do not go to the doctor, or (2) are deemed to have a negative health influence on the respondent, remain significant. We modified Table 4 and the text on P.8 to reflect the revised results.

P.8, line 175: To deconstruct these global effects of the social network, we examined the association of individual network metrics with the MSRS-R, adjusting for sex, age, marital status, and years of education (**Table 4**). None of the network structure metrics was significantly associated with MSRS-R score, consistent with the global assessment. Two network composition features were significantly associated with MSRS-R score: the percent of network members who (1) do not go to a doctor regularly, or (2) are deemed to have a negative health influence on the respondent. The strongest association was with the percent of network members who are deemed to have a negative health influence (beta=0.017±0.005, p=0.016).

3. In table 2 it would be useful to provide the actual percentages who smoke etc in addition to the medians IQR.

This is corrected to show the actual percentages with both a median value and the percentiles for the health habits data.

4. Could the authors also comment on how representative their sample was of the whole GEMS cohort would it be more likely that those who respond have a greater or lesser social network than those who did not?

The study population in this manuscript is representative of the overall cohort with respect to demographic profile. In this study, we sent out the social network questionnaire to all participants (as of October 2016) and set a data collection deadline 5 weeks later. We expect the participants who completed the social network questionnaire to be more engaged with the research study than those who did not. However, it is also possible that some participants were simply unavailable to complete the questionnaire during the time frame. It is very difficult to say whether those who completed the questionnaire have a different social network structure or composition than those who did not complete the questionnaire, simply because we do not have data from the latter.

1   August 14, 2018

2   **A Scalable Online Tool for Quantitative Social Network**

3   **Assessment Reveals Potentially Modifiable Social Environmental**

4   **Risks**

5   Amar Dhand[1,2,*], Charles C. White[3], Catherine Johnson[4], Zongqi Xia[5], Philip De Jager[4]

6

7   [1]Department of Neurology, Brigham and Women's Hospital, Harvard Medical School,

8   Boston, USA

9   [2]Network Science Institute, Northeastern University, Boston, USA.

10   [3]Broad Institute, Program in Medical and Population Genetics, Cambridge, USA.

11   [4]Multiple Sclerosis Center and the Center for Translational & Computational

12   Neuroimmunology, Department of Neurology, Columbia University Medical Center,

13   New York, USA.

14   [5]Department of Neurology, University of Pittsburgh, Pittsburgh, USA.

15

16   To whom correspondence should be addressed:

17   *Amar Dhand, M.D., D.Phil.

18   Brigham and Women's Hospital

19   75 Francis Street

20   Boston, MA 02115

21   617-732-5330; fax: 61-732-6083; Email: adhand@bwh.harvard.edu

22 **ABSTRACT**

23 Social networks are conduits of support, information, and health behavior flows.

24 Existing measures of social networks used in clinical research are typically summative

25 scales of social support or artificially truncated networks of ≤ 5 people. Here, we

26 introduce a quantitative social network assessment tool on a secure open-source web

27 platform, readily deployable in large-scale clinical studies. The tool maps an individual's

28 personal network including specific persons, their relationships to each other, and their

29 health habits. To demonstrate utility, we used the tool to measure the social networks of

30 1493 persons at risk of multiple sclerosis. We examined each person's social network in

31 relation to self-reported neurological disability. We found that the characteristics of

32 persons surrounding the participant, such as negative health behaviors, were strongly

33 associated with the individual's functional disability. This quantitative assessment

34 reveals key elements of individuals' social environments that could be targeted in

35 clinical trials.

36

37 **INTRODUCTION**

38   Social connectivity is known to impact health. Social isolation is a predictor of

39 mortality comparable to smoking, hypertension, and physical inactivity[1]. Social

40 enrichment has a strong positive effect on biological[2] and functional health outcomes[3,4].

41 Social connections are also potentially modifiable, making them ideal targets for

42 changing habits such as smoking, exercise, and diet[5].

43   Despite their promise in health, social networks are poorly understood in patient

44 populations and interventions aimed at networks are nascent. One main reason is a lack

45 of clear definition of the network surrounding a patient[6,7]. Traditional social network

46   metrics are actually summary indices of social support that query the total number of

47   social contacts, social resources available, and community engagement[8]. Multiple

48   clinical trials that have used such measures in patient populations have failed to

49   demonstrate a change in patient outcomes[9-11]. A more precise set of measures are

50   needed to map the specific people in the social system, one-by-one, and the kind of ties

51   between persons to clarify the network properties.

52   In this study, we introduce a social network assessment tool that quantifies

53   patients' personal network structure and health characteristics in a web-based, secure,

54   and scalable form. The tool is a survey adapted from a validated instrument, the General

55   Social Survey[12], and captures the structure of social ties and composition of

56   demographics and habits around the index patient. We demonstrate the utility of the

57   tool by quantifying the personal networks of 1493 individuals at-risk for multiple

58   sclerosis. The participants are enrolled in the Genes and Environment in Multiple

59   Sclerosis (GEMS) project, a prospective cohort study of people with first-degree family

60   history of MS[13]. The goal of the GEMS project is to identify novel genetic and

61   environmental risk factors, including the social environment. Prior work has shown that

62   asymptomatic MS family members who have a high burden of genetic and

63   environmental risk factors have evidence of diminished neurologic function [14]. Here, we

64   show a relationship in the GEMS cohort between social network metrics and

65   neurological disability. We demonstrate that quantifying social networks in large-scale

66   clinical studies offers an effective platform to identify previously unknown social

67   environment risk factors that could be potentially modifiable.

68

69

70   **RESULTS**

71   **Creating a scalable online tool to assess social networks**

72   We designed a HIPAA-compliant structured social network questionnaire

73   adapted primarily from the General Social Survey[12,15] (Supplementary **Methods 1**). The

74   schema of the data acquisition and potential use is presented in **Fig. 1**. The

75   questionnaire comprises approximately 48 questions with adaptation to responses. The

76   estimated completion time of the questionnaire is 10-15 minutes. The questionnaire

77   begins with three traditional name generators, in which participants named all people

78   with whom they had discussed important matters, socialized, or sought support in the

79   last 3 months. The number of people who could be named was not capped. Next,

80   participants answered questions that evaluate the connections between each pair of the

81   first 10 persons in the network, including the strength of ties in three levels (strangers,

82   weak, and strong). Finally, participants answered questions about the characteristics

83   and health habits of each of the first 10 persons in the network[7]. The online

84   questionnaire was hosted on the Research Electronic Data Capture (REDCap) server, a

85   secure web platform for administering questionnaires in clinical research[16]. A version of

86   the instrument is available for use in the REDCap Shared Library. Code to analyze and

87   visualize data created from the instrument is available on GitHub.

88   The assessment generated two main categories of network metrics, structure and

89   composition, based on graph theoretical statistics. Within the category of social network

90   structure, size is the number of individuals in the network, excluding the index

91   participant, or 'ego'. Density is a measure of connectivity of individuals in the network,

92   calculated as the sum of ties, excluding the ego's ties, divided by all possible ties[17].

93   Constraint is a more detailed version of density that quantifies the extent to which the

94   ego's connections are to individuals who are connected to one another. Effective size is

95   the number of non-redundant members in the network[18]. Maximum degree is the

96   highest number of ties by a network member, and mean degree is the average number of

97   ties by a network member. Equations for these measures are available in Supplementary

98   **Methods 2**.

99   Within the social network composition category, several metrics quantify the

100   ratio of member characteristics in the network. For instance, the percent kin is the

101   percent of individuals in the network who are family members. Standard deviation of

102   age represents the range of ages. The diversity of sex index is the mix of men and women

103   in the network, according to the index of qualitative variation[19], with a value of 1

104   indicating equal mix of men and women. The diversity of race is the mix of races

105   similarly calculated. Importantly, the questionnaire also queries the health behavior

106   environment around the participant by examining the percentage of the network

107   members with negative health habits, including smoking, sedentary lifestyle, not visiting

108   doctors regularly, and poor compliance of prescription medications. All compositional

109   variables were created to account for network size. Specifically, the number who fit a category

110   were divided by the total size to create the percentage.

111

112   **Demonstrating network quantification in a nation-wide cohort**

113   We assessed the social networks of 1493 GEMS participants from across the

114   United States (Supplementary **Fig. 1**), which represented 57% of the cohort as of

115   October 2016. In **Table 1**, we report the demographic and clinical information of the

116   cohort at the time of the study, separated into subgroups of asymptomatic participants

117   and participants with an MS diagnosis. Asymptomatic participants had a lower age on

118 average than participants with an MS diagnosis, consistent with the previously reported

119 demographics of the cohort[13].

120   The primary outcome measure of functional disability was the MSRS-R, a self-

121 reported outcome of functional disability validated for people with MS. The MSRS-R is a

122 brief questionnaire that correlates with traditional clinical instruments[20,21]. The eight

123 domains of MSRS-R include walking, using arms and hands, vision, speaking clearly,

124 swallowing, cognition, sensation, and bowel and bladder function, for a maximum score

125 of 32. In this cohort of primarily asymptomatic people at risk for MS, we chose MSRS-R

126 as an outcome measure because few alternative self-reported outcome measures have

127 the advantages of being concise and validated in early MS. As expected, the median

128 MSRS-R score was higher on average in the MS group than in the asymptomatic group.

129   To visualize each participant's social network structure, we plotted a montage of

130 all participants' networks, ranging from the smallest to the largest, with the strength of

131 each tie highlighted in color (**Fig 2**). The average network consisted of 8 people who

132 were densely linked (67% of all possible ties were present). Furthermore, an average

133 44% of all network members were kin, 38% were supportive of the index participant,

134 and there was a nearly equal mix of men and women (diversity score of 0.89 with 1

135 being an equal mixture of men and women). Race, on the other hand, was not varied

136 within networks with a diversity score of 0, indicating that most members in a

137 participant's network were of the same race. Weak ties, denoting those who are less

138 familiar with the participant, ranged from 20% to 67% depending on the measure. The

139 percent of individuals who were known for less than 6 years by the respondent was 20%

140 in asymptomatic persons and 12% in MS patients (p=0.001, Wilcoxon signed rank test),

141 suggesting a reduction in recent acquaintances in participants with a MS diagnosis.

142   Otherwise, differences in network structure and general network composition between

143   asymptomatic and MS participants were small and not significant (**Table 2**).

144   To visualize the milieu of health habits around the participant, we plotted a

145   montage of all participants' networks, ranging from the healthiest environment to the

146   least healthy (**Fig 3**). On average, the network composition with respect to health habits

147   skewed towards social environments in which most network members have healthy

148   habits. 17% of participants had personal networks in which all members were healthy.

149   On average, the percent of network members who do not exercise was 33%, and this was

150   the highest value out of the examined negative health habits. There was a weak negative

151   correlation between network size and the percentage of network members with

152   unhealthy habits (Pearson's correlation = -0.13±0.05, p<0.0001). Because we did not

153   detect differences in network composition with respect to healthy habits between

154   asymptomatic and MS participants, we were able to pursue joint analyses of these two

155   subgroups.

156   Having established the basic properties of our data, we examined the relationship

157   between network metrics and self-reported functional disability outcome. Given the

158   number of network metrics and to account for multiple testing burdens, we grouped the

159   network variables into structure and composition categories. We then used a

160   permutation based omnibus test to examine associations of these two groups of network

161   metrics with the MSRS-R. The observed distribution of P-values in the omnibus test was

162   greater than chance for network composition (p=<0.0001 all; p = 0.008 asymptomatic

163   subgroup; p=0.001 MS subgroup), but not for network structure (p=0.066 all; p=0.14

164   asymptomatic subgroup; p=0.25 MS subgroup) (**Table 3, Fig 4**). Thus, our global

165   assessments indicated that network composition, rather than network structure, was

166   associated with self-reported functional disability based on the MSRS-R scores (**Table**

167   **3**).

168   To deconstruct these global effects of the social network, we examined the

169   association of individual network metrics with the MSRS-R, adjusting for sex, age,

170   marital status, and years of education (**Table 4**). None of the network structure metrics

171   was significantly associated with MSRS-R score, consistent with the global assessment.

172   Two network composition features were significantly associated with MSRS-R score: the

173   percent of network members who (1) do not go to a doctor regularly, or (2) are deemed

174   to have a negative health influence on the respondent. The strongest association was

175   with the percent of network members who are deemed to have a negative health

176   influence (beta=0.017±0.005, p=0.016, linear regression).

177   In exploratory analyses, we examined the relationship between each individual's

178   Genetic & Environmental Risk Score (GERS) and her or his social network size. The

179   GERS is an aggregate estimate of an individual's MS risk based on validated genetic and

180   environmental susceptibility factors. We have previously reported that the GERS is

181   informative of MS risk beyond family history in the GEMS cohort of first-degree family

182   members.[13] Using the published GERS based on previously reported genetic and

183   environmental risk factor data available among a subset of the GEMS participants

184   (n=999 all, n=920 asymptomatic subgroup, n=79 MS subgroup), we noted an

185   association in linear regression between larger network size and increased GERS

186   (beta=0.82±0.19, p=2.43x10^{-5}, all) (Supplementary **Table 1**). This finding appears to be

187   driven by the larger network size of women participants relative to men. In a regression

188   analysis, network size is inversely related to male sex (beta=-1.87±0.42, p=8.71x10^{-6},

189   all). Among asymptomatic participants, both a history of mononucleosis

190   (beta=1.13±0.40, p=0.005) and a higher genetic risk score for MS susceptibility

191   (beta=0.65±0.24, p=0.006) were also associated with a larger network size in the linear

192   regression (Supplementary **Table 1**).

193

194   **DISCUSSION**

195   In this in-depth analysis of social networks in a clinical population, we

196   demonstrate the ease and utility of deploying our online questionnaire that evaluates an

197   individual's social network in a structured manner. In a few weeks and using only

198   electronic communication, we collected complete data on 1493 individual GEMS

199   participants. This large data set allowed us to pursue analyses in a statistically robust

200   manner and to produce highly significant results. These results represent an important

201   milestone in studies of MS and other neurologic conditions with a long prodromal

202   neurodegenerative phase: it provides investigators with the key data needed to support

203   power calculations and guide future study designs. In particular, we found that

204   asymptomatic family members at risk of MS have enough variance in our measure of

205   self-reported disability to yield strong association results with compositional but not

206   structural variables. Most prominently, the health habits of persons in their social

207   environment was strongly associated with the participant's self-reported neurological

208   dysfunction, and the percent of network members who have a negative health influence

209   had the strongest association with disability. While these results need to be validated,

210   they show (1) that studies of "at risk" individuals in which overt symptoms of a

211   neurologic disease have not yet become manifest are feasible and (2) that network

212   composition is an area that deserves further dissection in individuals at risk for MS and

213   perhaps for other neurodegenerative diseases.

214   Our assessment adds to a growing list of web-based personal network surveys

215   that translate the complexity and burdensome features of this type of questionnaire into

216   a more usable and scalable form[22]. Two examples in public health include EgoWeb

217   2.0[23], an open source software that may be used for motivational interviewing using

218   network graphics, and OpenEddi[24], a tool designed for interactive, tablet or mobile-

219   ready field collection of network data. Our tool is unique in that it is a HIPAA-compliant

220   data collection tool, able to be completed by patients without an interviewer, and has the

221   capability to handle large volumes of data from clinical populations using electronic

222   communications. The assessment also included questions customized for patients or at-

223   risk individuals with a focus on social support and health-related behaviors of network

224   members. These dimensions are critical for future planning of network interventions to

225   improve health and quality-of-life outcomes in clinical settings.

226   One mechanism that may explain some of our findings is the tendency of

227   individuals to associate with others who are similar to themselves, or homophily.

228   Similarity breeding social connection has been described in other social network

229   studies[25]. Race and ethnicity are the strongest linkage factors leading to homogenous

230   personal environments[25], and we found this in our study as well. However, there are

231   many examples of health behavior homophily. Children's social network composition is

232   significantly associated with several aspects of children's own health[26]. Latrine

233   ownership in rural India is correlated with latrine usage among social contacts, after

234   control of caste, education, and income[27]. An individual's weight is influenced by obesity

235   of spouses and same-sex social contacts[28], and incident type 2 diabetes is associated

236   with obesity in spouses[29]. Aspirin use is correlated with aspirin use among friends and

237   family[30]. Taken together, these findings point to core human behaviors that are shared

238   among like-minded social contacts, with eating and physical activity as major driving

239   forces for these effects.

240   Two more mechanisms that may explain the association of network members'

241   health habits and the participant's neurological disability are social contagion and

242   antecedent exposures. Social contagion is a type of social influence in which behavior in

243   one or many network members affects the behavior of the index participant. Detection

244   of this effect requires longitudinal data and network modeling, such as stochastic actor-

245   oriented or instrumental variable approaches, to understand the spread of behaviors

246   through social ties. For example, one study shows the spread of physical activity in 1

247   million users of a smartphone running app[31]. Antecedent exposures influencing both

248   parties may be another contributor. For example, rural environments with poor access

249   to medical services may influence the habits of all members of the network with regard

250   to seeking medical care. Finally, a combination of these factors may explain the

251   association of poor health habits in the network and a person's neurological disability.

252   The association between an individual's susceptibility for MS, as determined by

253   GERS, and social network size is a preliminary finding that requires further

254   investigation. This may be explained by the inclusion of sex as a component of GERS[13]

255   and prior observation that women tend to have larger social networks[15]. However, the

256   imbalance of men (19%) and women (81%) in this study potentially complicates the

257   interpretation. Another explanation is that larger network size reflects broader exposure

258   to infectious agents that are associated with MS susceptibility, such as history of

259   infectious mononucleosis[13]. Indeed, we observed a positive association between

260   mononucleosis and network size among asymptomatic participants. Finally, the role of

261 genetic factors in network size is provocative, but the effect is modest and needs further

262 investigation.

263 Our study has limitations. First, we were unable to establish causality and

264 directionality of the associations or the mechanisms of homophily in this cross-sectional

265 study. Within the GEMS platform, we are gathering longitudinal social network data.

266 Second, the primary outcome measure of neurological disability (MSRS-R) was skewed

267 towards low scores due to the larger proportion of self-reported asymptomatic

268 participants in the GEMS cohort who have low scores in this instrument. This could

269 reduce the precision of our analyses due to a floor effect. Further, the study may be

270 underpowered to compare asymptomatic and MS subgroups given the modest number

271 of the MS cases (i.e. familial MS). Larger studies of individuals with sporadic MS will

272 better answer whether social network variables influence disease worsening in MS.

273 Third, unmeasured confounders that influence report of social networks and functional

274 disability could have affected our findings. We attempted to address this limitation by

275 adjusting for major factors reported in the literature, including age, sex, and marital

276 status. Fourth, we ascertained social network metrics based on participants' self-report

277 of their social networks. While this approach may introduce unknown biases, prior work

278 reassuringly had shown self-reported personal networks of intimate contacts to be

279 accurate[32]. Finally, this study of the GEMS participants who were recruited through

280 advocacy groups, social media, and electronic communications, may not have broad

281 generalizability, because these participants are more socially engaged and better

282 educated than the general population. Future studies of more diverse populations and

283 other chronic neurological disorders will be critical.

284   The social environment is ubiquitous and important for understanding human

285   disease etiologies and outcomes. Social network features, in general, represent an

286   emerging group of metrics that inform aspects of health and disease, but are not

287   currently well captured by many biomedical research studies. We outline an approach of

288   quantitative social network analysis that is readily adaptable in clinical investigations.

289   The questionnaire that we have developed for quantifying social networks is available

290   through the open-source REDCap platform. In the empirical work described, we found

291   that the health behaviors of persons surrounding an individual at-risk for MS were

292   associated with the individual's own functional status. These results suggest that

293   interventions aimed at modulating network composition through education or

294   treatment of members in a social network holds the promise of a novel complementary

295   approach to managing MS onset and disease course.

296

297   **METHODS**

298   **Study design and participants.**

299   In a cross-sectional design, we invited GEMS participants to complete an online

300   questionnaire assessing social networks and current neurological disability in October

301   2016 (Supplementary **Methods 1**). The questionnaire was live for 6 weeks, with

302   reminders sent to non-responders. At the time, the GEMS cohort included 2,632 first-

303   degree family members from across the United States, recruited using patient advocacy

304   groups, social media, and word of mouth [13]. The inclusion criteria were: being 18 to 50

305   years of age at enrollment, and having at least one first-degree relative with a diagnosis

306   of MS (e.g., parent, full-sibling, or child). While asymptomatic family members who are

307   at risk for MS represent the main focus of the GEMS project, we also recruited family

308   members who already have a MS diagnosis for comparison in this cross-sectional study.

309   MS cases were confirmed by review of medical records. The institutional review boards

310   of all participating sites (Partners HealthCare, National Institutes of Health, and

311   University of Pittsburgh) approved the study. All participants provided written informed

312   consent.

313

314   **Statistical methods.**

315   To compare demographic characteristics between asymptomatic participants and

316   confirmed MS cases, we performed a t-test for age, chi-squared tests for dichotomous

317   variables of sex, marital status, and living alone, as well as non-parametric Wilcoxon

318   rank-sum tests for years of education and MSRS-R. Similarly, we performed non-

319   parametric Wilcoxon rank-sum tests to compare network metrics between

320   asymptomatic participants and participants with MS diagnosis.

321   To assess the association with MSRS-R score, we performed a linear regression

322   for each network variable, adjusting for age, sex, and marital status. In this analysis,

323   MSRS-R was modeled as the dependent variable, and each network characteristic as the

324   independent variable. Within each network metrics category (structure and

325   composition), we calculated the false discovery rate to adjust for multiple testing. To

326   examine any potential bias due to non-normal distributions, we performed a sensitivity

327   analysis applying non-parametric spearman correlation tests.

328   To examine the hypothesis that as a category, social network variables were

329   associated with the MSRS-R score, we performed an empirical omnibus test. In the first

330   stage of this analysis, we calculated the p-values of association between each network

331   variable and MSRS-R score using linear regression as described above. In the second

332   stage, we used a Fisher's meta-analysis to combine these p-values and calculate a chi-

333   squared statistic. We then compared this chi-square statistic to an empirical distribution

334   of chi-squared statistics as generated by 10,000 random permutations. By permuting

335   the MSRS-R score, we maintained the correlation structure of the network variables.

336   The empirical omnibus p-value was then calculated as the number of times that the chi-

337   squared statistic from the 10,000 permutations was greater than the true chi-squared

338   statistic, divided by the total number of permutations. To generate a quantile-quantile

339   plot, we plotted the observed -log10(p-value) of each pair of association between a

340   network variable and MSRS-R score against the expected -log10(p-value). The 90th and

341   95th empirical confidence intervals were determined using empirical p-values as

342   generated by the 10,000 permutations. We performed the omnibus test in all

343   participants as well as in the subset of asymptomatic participants and the subset of

344   participants with MS diagnosis.

345   In exploratory analyses, we assessed the relationship of GERS (a published

346   estimate of MS risk based on an individual's known genetic burden and environmental

347   exposures for MS susceptibility) and social network metrics. Here, we performed linear

348   regressions, adjusting for age, modeling network size as the dependent variable, and the

349   GERS (and its components: history of infectious mononucleosis, sex, smoking status;

350   environmental risk score; genetic risk score) as the independent variables. All analyses

351   were performed in R version 3.2[33]. All statistical tests were two-sided. Given the

352   exploratory nature of the analysis and data, power calculations were not performed

353   prior to analysis. Permutations and nonparametric tests were used to avoid bias due to

354   any non-normal data or unequal variances between groups, as necessary.

355

Code availability.

An updated version of the instrument called "Personal Network Survey for Clinical

Research" is available in the REDCap Shared Library. We have also uploaded a

comprehensive R codebase for researchers who use the instrument to analyze and

visualize their data available at: https://github.com/AmarDhand/PersonalNetworks. R

code used specifically for this project can be made available upon request.

Data availability.

The data used in this study is freely available as a supplement to this manuscript

(Supplementary **Database 1**).

AUTHOR CONTRIBUTIONS

A.D., C.W., Z.X., and P.L.D conceived the study. A.D., C.W., and C.J. collected the data.

A.D., C.W. Z.X., C.J., and P.L.D performed data analysis. A.D., C.W. Z.X., C.J., and

P.L.D wrote the manuscript.

ADDITIONAL INFORMATION

Supplementary information is available in the online version of the paper.

Completing interests. The authors have declared that no conflict of interest exists.

380

REFERENCES

382   1.   Holt-Lunstad, J., Smith, T. B. & Layton, J. B. Social relationships and mortality

383   risk: a meta-analytic review. *Plos Med.* **7,** e1000316 (2010).

384   2.   Yang, Y. C. *et al.* Social relationships and physiological determinants of longevity

385   across the human life span. *Proc. Natl. Acad. Sci. USA* 113, 578–583 (2016).

386   3.   Fratiglioni, L., Wang, H. X., Ericsson, K., Maytan, M. & Winblad, B. Influence of

387   social network on occurrence of dementia: a community-based longitudinal study.

388   *Lancet.* **355,** 1315–1319 (2000).

389   4.   Glass, T. A., Matchar, D. B., Belyea, M. & Feussner, J. R. Impact of social support

390   on outcome in first stroke. *Stroke.* **24,** 64–70 (1993).

391   5.   Valente, T. W. *Social Networks and Health.* (Oxford University Press, New York,

392   2010).

393   6.   Legg, L. A. *et al.* Non-pharmacological interventions for caregivers of stroke

394   survivors. *Cochrane Database Syst. Rev.* CD008179 (2011).

395   7.   Dhand, A., Luke, D. A., Lang, C. E. & Lee, J. M. Social networks and neurological

396   illness. *Nat. Rev. Neurol.* **12,** 605–612 (2016).

397   8.   Smith, K. P. & Christakis, N. A. Social networks and health. *Annu. Rev. Sociol.* **34,**

398   405–429 (2008).

399   9.   Glass, T. A. *et al.* The Families In Recovery From Stroke Trial (FIRST): primary

400   study results. *Psychosom. Med.* **66,** 889–897 (2004).

401   10.   Berkman, L. F. *et al.* Effects of treating depression and low perceived social

402   support on clinical events after myocardial infarction: the Enhancing Recovery in

403   Coronary Heart Disease Patients (ENRICHD) Randomized Trial. *JAMA.* **289,**

404   3106–3116 (2003).

405   11.   Forster, A. *et al.* A structured training programme for caregivers of inpatients

406   after stroke (TRACS): a cluster randomised controlled trial and cost-effectiveness

407   analysis. *Lancet.* **382,** 2069–2076 (2013).

408   12.   Burt, R. S. Network items and the general social survey. *Social Networks.* **6,** 293–

409   339 (1984).

410   13.   Xia, Z. *et al.* Genes and Environment in Multiple Sclerosis project: A platform to

411   investigate multiple sclerosis risk. *Ann. Neurol.* **79,** 178–189 (2016).

412   14.   Xia, Z. *et al.* Assessment of Early Evidence of Multiple Sclerosis in a Prospective

413   Study of Asymptomatic High-Risk Family Members. *JAMA Neurol.* **74,** 293–300

414   (2017).

415   15.   O'Malley, A. J., Arbesman, S., Steiger, D. M., Fowler, J. H. & Christakis, N. A.

416   Egocentric social network structure, health, and pro-social behaviors in a national

417   panel study of Americans. *PLoS ONE.* **7,** e36250 (2012).

418   16.   Harris, P. A. *et al.* Research electronic data capture (REDCap)--a metadata-driven

419   methodology and workflow process for providing translational research

420   informatics support. *J. Biomed. Inform.* **42,** 377–381 (2009).

421   17.   Wasserman, S. & Faust, K. *Social Network Analysis.* (Cambridge University

422   Press, Cambridge, 2009).

423   18.   Burt, R. S. Structural Holes and Good Ideas. *American Journal of Sociology.* **110,**

424   349–399 (2004).

425   19.   Agresti, A. & Agresti, B. F. Statistical Analysis of Qualitative Variation.

426   *Sociological Methodology.* **9,** 204–237 (1978).

427   20.   Wicks, P., Vaughan, T. E. & Massagli, M. P. The multiple sclerosis rating scale,

| 428 | | revised (MSRS-R): development, refinement, and psychometric validation using |

an online community. *Health Qual. Life Outcomes.* **10,** 70 (2012).

21. Bove, R. *et al.* Complex relation of HLA-DRB1*1501, age at menarche, and age at multiple sclerosis onset. *Neurol. Genet.* **2,** e88 (2016).

22. Vehovar, V., Lozar Manfreda, K., Koren, G. & Hlebec, V. Measuring ego-centered social networks on the web: Questionnaire design issues. *Social Networks.* **30,** 213–222 (2008).

23. Kennedy, D. P. *et al.* A computer-assisted motivational social network intervention to reduce alcohol, drug and HIV risk behaviors among Housing First residents. *Addict. Sci. Clin. Pract.* **11,** 4–17 (2016).

24. Eddens, K. S., Fagan, J. M. & Collins, T. An Interactive, Mobile-Based Tool for Personal Social Network Data Collection and Visualization Among a Geographically Isolated and Socioeconomically Disadvantaged Population: Early-Stage Feasibility Study With Qualitative User Feedback. *JMIR Res. Protoc.* **6,** e124–17 (2017).

25. McPherson, M., Smith-Lovin, L. & Cook, J. M. Birds of a Feather: Homophily in Social Networks. *Annu. Rev. Sociol.* **27,** 415–444 (2001).

26. Kennedy-Hendricks, A. *et al.* Health implications of social networks for children living in public housing. *Health Place.* **36,** 145–151 (2015).

27. Shakya, H. B., Christakis, N. A. & Fowler, J. H. Social network predictors of latrine ownership. *Soc. Sci. Med.* **125,** 129–138 (2015).

28. Christakis, N. A. & Fowler, J. H. The spread of obesity in a large social network over 32 years. *N. Engl. J. Med.* **357,** 370–379 (2007).

29. Raghavan, S. *et al.* Incident Type 2 Diabetes Risk is Influenced by Obesity and

452   Diabetes in Social Contacts: a Social Network Analysis. *J. Gen. Intern. Med.* **31,** 1–

453   7 (2016).

454   30.   Strully, K. W. *et al.* Aspirin use and cardiovascular events in social networks. *Soc.*

455   *Sci. Med.* **74,** 1125–1129 (2012).

456   31.   Aral, S. & Nicolaides, C. Exercise contagion in a global social network. *Nat.*

457   *Commun.* **8,** 14753 (2017).

458   32.   Green, H. D., Hoover, M. A., Wagner, G. J., Ryan, G. W. & Ssegujja, E. Measuring

459   Agreement between Egos and Alters: Understanding Informant Accuracy in

460   Personal Network Studies. *Field Methods* **26,** 126–140 (2013).

461   33.   R Core Team. R: A language and environment for statistical computing. (2016).

462

463 **TABLES**

464

Table 1: Demographics and Clinical Characteristics of the Participants			
Characteristic	Asymptomatic n=1378	MS n=115	P-value[a]
Age, mean (SD), y	37.85 (8.34)	43.14 (7.60)	<0.001
Male sex, No. (%)	269 (19.5)	19 (16.5)	0.51
Years of education, median [IQR]	16 [16, 18]	16 [15, 18]	0.18
Married, No. (%)	914 (66.7)	86 (76.1)	0.051
Living alone, No. (%)	198 (13.4)	12 (10.4)	0.45
Age of onset of MS symptoms, mean (sd)	NA	30.50 (8.70)	NA
Age of diagnosis of MS, mean (sd)	NA	34.36 (7.74)	NA
MSRS-R, median [IQR]	1.00 [1.00, 2.00]	7.00 [3.00, 11.00]	<0.001

465 [a] P values calculated using t-test for age; chi-squared test for female sex, married, and living

466 alone; and Wilcoxon signed rank test for years of education and MS rating scale score-Revised

467 (MSRS-R).

468

469

Table 2: Network Characteristics

Characteristic	Asymptomatic n=1378	MS n=115	P-value[a]
Network Structure[b]			
Size, median [IQR]	8.00 [6.00, 12.00]	8.00 [5.00, 11.50]	0.130
Density, median [IQR]	67.00 [50.00, 89.00]	69.00 [53.00, 90.00]	0.170
Constraint, median [IQR]	44.00 [37.72, 53.03]	44.71 [38.19, 56.17]	0.315
Effective Size, median [IQR]	4.00 [2.80, 5.25]	3.55 [2.50, 5.07]	0.053
Maximum Degree, median [IQR]	5.00 [4.00, 7.00]	5.00 [4.00, 8.00]	0.987
Mean Degree, median [IQR]	4.00 [2.80, 5.00]	4.00 [2.50, 5.40]	0.493
Network Composition– General[c]			
Percent kin, median [IQR]	43 [30, 62]	50 [33, 67]	0.205
Percent who are supportive, median [IQR]	38 [25, 50]	40 [21, 50]	0.561
Standard deviation of age, median [IQR]	12.76 [10.04, 15.38]	12.98 [10.54, 16.89]	0.161
Diversity of sex, median [IQR]	0.89 [0.64, 0.96]	0.82 [0.64, 0.96]	0.108
Diversity of race, median, Percentile {90th,95th,99th,100th}[d]	0 {0.44,0.55,0.72,1.20}	0 {0.41,0.59,0.77,0.77}	0.046
Percent contacted weekly or less often, median [IQR]	67 [50, 80]	67 [45, 80]	0.896

Percent who have been known for less than 6 years, median [IQR]	20 [0, 43]	12 [0, 33]	0.001
Percent who live more than 15 miles away, median [IQR]	33 [17, 50]	33 [20, 56]	0.514

Network Composition–Health Habits[e]

Percent who smoke, median [IQR]	0 [0, 20]	0 [0, 40]	0.164
Percent who do not exercise, median [IQR]	33 [14, 54]	25 [10, 50]	0.068
Percent who do not take medications regularly, median, Percentile {90th,95th,99th,100th}	0 {0, 14, 33, 100}	0 {0, 17, 24, 50}	0.709
Percent who do not go to doctor's appointments, median, Percentile {90th,95th,99th,100th}	0 {0, 12, 25, 100}	0 {0, 15, 48, 100}	0.314
Percent who have a negative influence on health, median, Percentile {90th,95th,99th,100th}	0 {29, 46, 71, 100}	0 {20, 33, 78, 100}	0.150

470   [a] P-values calculated using Wilcoxon signed rank test.

471   [b] Network structure is quantified into graph theoretic statistics. See definitions in Methods.

472   [c] Network composition–General is the range of characteristics of people around the participant.

473   See definitions in Methods.

474   [d] Percentile are used to better understand the right-skewed distribution of the variables of race

475   and certain health habits .

476   [e] Network composition–Health Behavior is the range of health habits of people around the

477   participant.

478

Table 3: Relationship of the Composite Categories of Network Variables to MSRS in all participants

Variable Category	Number of Variables	Top variable	Top variable P-value	Top variable FDR[a] value	Composite P-value[b]
Structure	6	Total Size	0.025	0.133	0.066
Composition	13	Percent who do not go to doctor's appointments	7.4×10^{-8}	9.6×10^{-7}	<0.0001

479 [a] FDR, False discovery rate
480 [b] Permutation based omnibus test is described in the methods.
481

482

Table 4: Relationship of Individual Network Variables to MSRS-R

Variable	Beta	Standard Error	Adjusted P-value[a]	FDR[b]
Network Structure				
Size	-0.025	0.013	0.052	0.197
Density	0.007	0.365	0.985	0.985
Constraint	0.004	0.007	0.537	0.729
Effective Size	-0.035	0.05	0.487	0.712
Maximum Degree	-0.041	0.044	0.347	0.564
Mean Degree	0.003	0.052	0.958	0.985
Network Composition–General				
Percent kin	0.001	0.004	0.769	0.876
Percent who are supportive	-0.005	0.004	0.198	0.47
Standard deviation of age	-0.006	0.017	0.701	0.876
Diversity of sex	-0.332	0.359	0.356	0.564
Diversity of race	0.686	0.423	0.105	0.333
Percent contacted weekly or less often	-0.009	0.004	0.023	0.147
Percent who have been known for less than 6 years	0.001	0.004	0.784	0.876
Percent who live more than 15 miles away	-0.003	0.004	0.346	0.564
Network Composition–Health				
Percent who smoke	0.006	0.003	0.045	0.197
Percent who do not exercise	-0.003	0.003	0.296	0.564
Percent who do not take medications	0.018	0.012	0.123	0.334

regularly				
Percent who do not go to doctor's appointments	0.045	0.015	0.002	0.023
Percent who have a negative influence on health	0.017	0.005	0.001	0.016

483 [a] Adjusted for potential confounders, sex, age, marital status, and years of education via linear

484 regression as described in the methods.

485 [b] FDR is false discovery rate, controlling for multiple testing.

486

FIGURE LEGENDS

Fig. 1. Overview of data collection, analysis, and interventions. This flow-chart shows the social network data acquisition, identification of modifiable elements in the social environment, and potential intervention strategies.

Fig 2: Structure of participants' personal social network. Each small network has a black circle that represents the participant, who is surrounded by white circles who are the network members. The lines connecting the circles are red if the relationship is strong and blue if the relationship is weak. Networks are arranged from the smallest (top left) to the largest (bottom right).

Fig 3: Health habits in participants' personal social network. In each network, a black circle is the participant, a white circle is a healthy social contact, and a red dot is an unhealthy social contact. Unhealthiness is defined as someone who does any of the following: smokes, does not exercise, does not visit doctors regularly, or not compliant with prescription medications. Networks are arranged from least negative health influence (top left) to most negative health influence (bottom right).

Fig 4: Comparison of expected versus observed regression results. Quantile-quantile plot of expected versus observed P values of composite network structure and network composition metrics in relation to neurological function and disability in the full cohort (A, B) and subgroups of asymptomatic (C, D) and MS participants (E, F). The expected P values ($-\log_{10}$[P value]) are shown on the x-axis, and the observed P values ($-\log_{10}$[P value]) are shown on the y-axis. The dark gray area indicate the confidence

512   interval ranges as generated by chance at a threshold of P= 0.10, and the light grey is for

513   P= 0.05. The observed values for composition, and not structure, are outside of the grey

514   areas, suggesting that composition is associated with the MSRS-R score beyond chance

515   after accounting for multiple testing burden and correlation structure of the

516   composition variables.

517

518