POPULATION HEALTH MANAGEMENT
Volume 21, Number 4, 2018
Mary Ann Liebert, Inc.
DOI: 10.1089/pop.2017.0097

How to Measure Population Health: An Exploration Toward an Integration of Valid and Reliable Instruments

Roy J.P. Hendrikx, MSc,[1] Marieke D. Spreeuwenberg, PhD,[2,3] Hanneke W. Drewes, PhD,[4] Dirk Ruwaard, PhD,[2] and Caroline A. Baan, PhD[1,4]

Abstract

Population health management initiatives are introduced to transform health and community services by implementing interventions that combine various services and address the continuum of health and well-being of populations. Insight is required into a population's health to evaluate implementation of these initiatives. This study aims to determine the performance of commonly used instruments for measuring a population's experienced health and explores the assessed concepts of population health. Survey-based Short Form 12, version 2 (SF12, health status), Patient Activation Measure 13 (PAM13), and Kessler 10 (K10, psychological distress) data of 3120 respondents was used. Floor/ceiling effects were studied using descriptive statistics. Validity was assessed using factor and discriminant analyses, and reliability was assessed using Cronbach α. Finally, to study covered concepts, exploratory factor analyses (EFAs) were conducted, which included additional surveyed characteristics. The SF12 and PAM13 sum scores showed acceptable averages and distributions, while results of the K10 indicated a floor effect. SF12 and K10 measured their expected constructs, while PAM13 did not. The EFA of PAM13 displayed 1 instead of the expected 4 constructs. Reliability was good for all instruments (α 0.89–0.93). The overall EFA identified 4 concepts: mental, physical ability, lifestyle, and self-management. SF12 and PAM13, combined with lifestyle characteristics, are shown to provide insightful information to measure the physical, mental, lifestyle, and self-management concepts of population health. Future research should include additional instruments that cover new aspects introduced by recent definitions of health.

Keywords: population health, population management, Triple Aim, evaluation

Background

Population (health) management (PM) initiatives are a response to the pressure put on health care systems by aging populations and new expensive technological possibilities.[1] PM aims to address this burden by focusing on a defined, often general, population's complete continuum of health and well-being, and integrating care across multiple care domains.[1] In order to be successful, PM should simultaneously improve the health of the population and the quality of care, while reducing cost growth (Triple Aim).[2] Therefore, to assess the implementation of PM initiatives, insight is needed into population health.[3] A great number of potential instruments and measures exist in order to gain this insight,[4] but for many, knowledge regarding performance in a general population is limited.

[1]Tranzo Scientific Center for Care and Welfare, Tilburg School of Social and Behavioral Sciences, Tilburg University, Tilburg, the Netherlands.
[2]Department of Health Services Research, Faculty of Health, Medicine and Life Sciences, CAPHRI–Care and Public Health Research Institute, Maastricht University, Maastricht, the Netherlands.
[3]Research Centre for Technology in Care, Zuyd University of Applied Sciences, Heerlen, the Netherlands.
[4]Department for Quality of Care and Health Economics, Center for Nutrition, Prevention and Health Services, National Institute for Public Health and the Environment, Bilthoven, the Netherlands.

This study was presented as a poster presentation at the International Conference on Integrated Care 2017, held May 8–10, 2017 in Dublin, Ireland.

Population health is not an easy concept to measure, as the concept and its considered constructs have shown little consensus over the years.[5] Kindig and Stoddart provide the most commonly used definition, stating that population health is ''the health outcomes of a group of individuals, including the distribution of such outcomes within the group.''[6] Additionally, for the concept of health itself, the World Health Organization created a definition in 1948 that is still used today: ''health is a state of complete physical, mental and social well-being and not merely the absence of disease or infirmity.''[7] More recently, the definition of health has been broadened and focuses more on the individual's ''ability to adapt.'' This led to the addition of constructs such as self-management, functioning, and (social) participation.[8,9]

Various measures, both objective and subjective, have been developed to assess (constructs of) health. Measures such as mortality and disease-specific prevalence statistics often are suggested and used for assessing objective health.[3,10] These measures are proven and still used, but there has been an increased focus on including experienced health.[11] For experienced health, the number and quality of survey instruments varies greatly per construct, ranging from general quality of life, which has many instruments available (eg, Short Form 12, EuroQol 5 Dimensions), to self-management (eg, Patient Activation Measure 13 [PAM13], Self-Management Ability Scale). For these instruments to be useful for PM initiatives, in addition to the validity and reliability requirements that apply to all instruments,[12] they need to conform to 2 criteria.[13] First, instruments should create only small amounts of missing data. Even though there are statistical methods to deal with missing data,[14] low response rates can indicate a lack of an instrument's applicability within the studied population. Second, an instrument should provide a wide range of responses across its items, providing room for improvement or degradation when evaluating PM initiatives. Additionally, survey length is an important consideration when measuring a complex concept such as experienced health. Longer surveys are associated with lower response rates,[15] and thus every instrument should measure distinct constructs to prevent redundancy.

In the Netherlands, the National Monitor Population Management (NMP) applied a potential set of instruments to evaluate population health within 9 PM initiatives. set included health-related characteristics (eg, body mass index [BMI]) combined with the Short Form 12 version 2 (SF12, health status), Kessler 10 (K10; psychological distress), and PAM13 (self-management).[16] These were selected based on expert suggestions and validation studies.[17–19] However, there is limited information on whether SF12, PAM13, and K10 meet the aforementioned criteria, making it unclear whether these instruments, individually or combined with other variables, can be used to measure a population's experienced health.

This study will examine the usability of common instruments for the evaluation of experienced health in the general population. First, the performance of each instrument (SF-12, K10, and PAM13) regarding missing data, distribution of scores, and the ability to differentiate between (sub)populations, as well as the validity and reliability in this setting, will be analyzed. Second, this study performs exploratory analyses to determine which constructs of population health (eg, physical, mental, social) these instruments measure when combined with other health-related characteristics.

Methods

Ethics approval and consent to participate

The Psychological Ethics Committee of the Tilburg University (EC-2014.39) approved this study.

Study population

The population consisted of citizens living in areas served by 9 Dutch PM initiatives. For analysis, data collected by the NMP survey (Psychological Ethics Committee number: EC-2014.39), conducted in December of 2014 and January of 2015, were used. A random sample was drawn from each initiative, comprising 600 insured adults (≥18 years old) who did not receive a previous survey from insurers and did not have a general physician registration fee in the past year. In addition, a national random sample of 1200 people outside these initiatives received a survey, bringing the total to 6600 surveys. Participants were invited by mail, asking if they were willing to fill out the survey online or on paper. Those who did not respond received 2 reminders by mail.[16]

Survey instruments

Preferred instruments were selected using input from national experts and literature. This process is described in more detail in Supplementary File S1 (Supplementary Data are available online at www/liebertpub.com/pop).

Health-related characteristics. In addition to demographic and socioeconomic characteristics (sex, age, education, employment, and origin), the survey covered several health-related characteristics. These included disability, physical exercise, alcohol use, and smoking status. Furthermore, the height and weight provided were combined to calculate BMI, and Chew et al's Set of Brief Screening Questions was used to assess health literacy. Care use was determined using the summed number of reported visits to care professionals.

SF12. The SF12 is a generic health status instrument consisting of 12 items.[20] This instrument was selected because it is used worldwide, in many different populations, and is able to produce both a physical component score (PCS) and a mental component score (MCS). For this study, the Dutch version was used.[21] The 12 items were scored and processed using the proprietary Scoring Software provided by QualiMetrics Inc. (Sacramento, CA). In short, this assigns all possible answers to each item its own weight and these are used to produce raw scores that are transformed to a 0–100 scale, in which a higher score means better health.[20]

PAM13. PAM13 scale consists of 4 stages: the patient believing his/her role is important, having the confidence and knowledge to take action, actually taking action, and staying the course even under stress.[22,23] Thirteen questions are weighted and combined to create a final score on a 0–100 scale. PAM13 was included because of its

comprehensive coverage of the concept as well as the positive association with various health-related behaviors (eg, lifestyle, medication adherence).[19] The Dutch version was used.[19]

K10. K10 is a 10-question scale that was initially included as a screening instrument for depression.[18] However, its use may be broader as it could potentially measure physiological distress in populations.[24] The scores, ranging from 1 to 5 (Likert-scale) per item, are added to create a sum score. This sum score ranges from 10, indicating no distress, to 50, indicating severe distress.[25] The Dutch version of the K10 scale was used in this study.[26]

Analysis

All analyses were performed using SPSS 22 (IBM Corporation, Armonk, NY) and R Studio Version 0.99.441 for Windows (RStudio Inc., Boston, MA).

First, item and overall instrument response rates were analyzed. Additionally, for each instrument, respondents with 1 or more missing values were compared to respondents who supplied a complete reply for that instrument. If groups differed, then it was shown that missing values were not likely to be Missing-Completely-At-Random and would need to be imputed using the multiple imputation by chained equation procedure.[27] This procedure created 5 complete data sets in which missing values were filled in (imputed) based on their correlation (minimal correlation = 0.3) with other variables. Results from the imputed data sets were combined using Rubin's rule.[27] All instrument sum scores were calculated in each of the 5 imputed data sets, except those of SF12, PCS, and MCS, because of the required and recommended use of the proprietary Scoring Software, which has its own algorithm for missing data. Further analyses were performed, where applicable, on both the original data set (listwise deletion) as well as the imputed data sets.

Second, overall sum score descriptives, including maximum, minimum, average, median, standard deviation, and variance, were determined as well as proportions at the highest and lowest end of each instrument's items, and sum scores to check for floor and/or ceiling effects. Additionally, box plots using quartiles visualized the spread of sum scores for each instrument.

Third, construct, convergent, and discriminant validity were studied separately for each instrument. To determine if the expected constructs came forward from the data (construct validity), a confirmatory factor analysis (CFA) was executed. For SF12 this was based on a model entailing 2 uncorrelated latent factors, ''mental'' and ''physical,'' to which all 12 items could contribute.[28] K10's model consisted of 1 factor, ''distress,'' that covered all items.[18] Four factors were included in the CFA of PAM13: ''believes'' (item 1 and 2), ''confidence'' (item 3 through 9), ''action'' (item 10 through 12), and ''stress'' (item 13).[22] Sampling adequacy was assessed using the Kaiser-Meyer-Olkin (KMO) test (should be >0.5) and Barlett's test for sphericity (should be $P < 0.05$). The fit was assessed using several goodness-of-fit indexes, including the root mean squared error of approximation (RMSEA). If the CFA resulted in a bad fit or provided inconclusive results, then an additional data-driven exploratory factor analysis (EFA) was performed to identify alternative constructs. This EFA was based on maximum likelihood analysis and varimax rotation. Sensitivity analyses were performed for both the CFAs and the EFAs by splitting the sample population by age: younger and older than age 65. Convergent validity, the amount to which related constructs are actually related, was assessed using corrected item-total correlations (>0.3 was acceptable). Discriminant validity was tested by seeing if instruments were able to discriminate between the populations younger and older than age 65, and non–highly educated and highly educated participants. This was assessed using independent t tests.

Fourth, reliability of all 3 instruments was assessed separately, mainly by internal consistency as calculated by Cronbach α.

Finally, to explore which health concepts were covered by the survey as a whole (the second research question), an EFA was performed including the sum scores of the SF12 (PCS and MCS), K10, and PAM13, as well as the other gathered health-related characteristics. This EFA was based on a maximum likelihood analysis and oblique rotation (direct oblimin) as it was assumed that the factors would correlate. Additionally, bivariate Pearson correlations between all outcome measures were checked. These were expected to show at least some consistency, as all instruments are health-related. These steps were repeated with participants younger and older than age 65.

Results

Results reported in the text are based on the analyses performed using listwise deletion. Multiple imputation-based results can be found in Supplementary File S2.

Description of sample

In total, 6600 surveys were sent out, 3120 of which were completed and returned (Table 1). The study population was less than 50% male and consisted of mostly Dutch natives; almost one third were older than age 65. Compared to the

Table 1. Characteristics of Sample Population

Population	*Study population*	*Dutch population*[1]
Surveys sent	6600	-
Surveys returned	3120	-
Response rate (%)	47.3	-
Sex (% male)	46.2	49,2
Age (% 65+)	29.3	15.9
Education (% highly educated)	26.7	23.6
Origin (% native)	86.7	77.9
Employed (% paid job)	49.6	63.2
Disabled (%)	4.1	3.9
Weight (% overweight, BMI ≥25)	56.3	48.3
Alcohol use (% excessive alcohol users)	4.2	8.4
Smoking (% smokers)	17.7	22.7
Inadequate health literacy (% ≤2 score Chewet al's Set of Brief Screening Questions)	6.1	-

[1]Based on CBS Statline descriptives.[29–31]
BMI, body mass index.

general Dutch population (Table 1), most characteristics were similar, but the proportion of people older than age 65 in the study population was almost twice that of the general Dutch population.

Response rates and missing values

The response rates to SF12 were lower and had a greater dispersion when compared to PAM13 and K10 (Supplementary File S3). Items 4 through 7 of SF12, relating to work, had the lowest response rates (range 87.0%–89.4%). The range of response rates of PAM13 and K10 were 97.3%–98.0% and 97.9–98.3%, respectively. The reply ''not applicable'' was used in all items of PAM13, ranging from 2.4% to 32.6%. Items 4 (32.6%), 8 (25.4%), and 9 (28.1%) had the highest percentage of respondents indicating ''not applicable.''

SF12 had the highest percentage of respondents with 1 or more missing values (31.2%). Less than a quarter of PAM13 (7.5%) and K10 (5.1%) had missing values (see Supplementary File S2). The employment rate of respondents was significantly lower for the 1-or-more-missing group for SF12, as was health literacy. When comparing the no-missing groups with the 1-or-more-missing groups, the latter group was significantly older in all instruments. For PAM13, the no-missing group was significantly more highly educated, while alcohol use was higher for this group for both PAM13 and K10. These results showed that the no-missing group and 1-or-more-missing groups differed, warranting the use of multiple imputation. Results of the analyses based on multiple imputed data can be seen in the tables of Supplementary File S2.

Ceiling/floor effect

SF12 (Supplementary File S3) had low percentages (1.3%–8.2%) of respondents with the lowest score, while 11.3% to 66.2% of respondents chose the highest score. Two thirds of respondents picked the most positive answer out of the possible 5 answers for items 2 and 3. For items 6, 7, 8, and 12, this was more than 40%. PAM13 had a similar range, between 1.3% and 3.0% of respondents, with the lowest score and a higher number (14.2%–37.5%) of respondents selecting the highest score. The not-applicable option, which is considered a missing value in PAM13 syntax, ranged between 2.4% and 28.1% per item. More than a quarter of respondents used this option in items 4, 8, and 9. K10 had the lowest level of respondents selecting the lowest score (0.4%–2.3%), and had high percentages at the other end of the scale: per item, between 36.4% and 72.8% of respondents indicated the highest score, with most items well above 50%.

The sum scores of SF12 and PAM13 (Table 2) showed a good spread, while K10 seemed to present a floor effect (lower is better for this instrument). This can be seen in the average scores, median, quartiles, and most clearly in the box plots (Supplementary File S4). Each instrument showed its full or almost full range (minimum to maximum) of possible values.

Validity

The CFAs yielded inconclusive results regarding construct validity (Table 3). SF12 did not converge using the complete cases as well as the imputed data. Its model, which lets all variables load freely on 2 identical factors, prevented the factors from being identified. PAM13 and K10 converged in both data sets and provided comparable results. When looking at the goodness-of-fit analyses from the CFAs that converged, they seemed to be contradictory (Table 3). For PAM13 and K10 the chi-square results showed *P* values below 0.001, indicating the original constructs were not found in this data set. The RMSEA confirmed this, but the other goodness-of-fit indexes all show a moderate-to-good fit for each instrument. In the subsequent EFAs (Supplementary File S5), all KMO tests were satisfactory (SF12 = 0.917, PAM13 = 0.891, K10 = 0.932) and each sphericity test was significant ($P \leq 0.001$). The scree plots provided a clear number of factors for each instrument. SF12 showed 2 factors, clearly indicating the expected physical and mental factor. K10 also was in accordance with its intended model. Repeating these analyses among participants older and younger than age 65 did not yield any different results. The EFA of PAM13 resulted in a single factor, which differed from the expected 4 factors. The CFA and the EFA contradict on several points; this is probably the result of the more strict nature of CFA. A CFA requires zero factor loading on other factors, while the EFA is more forgiving in this regard.[32]

Convergent validity was acceptable across all instruments, as corrected item-total correlations were all above

Table 2. Descriptives of the Sum Scores of the Short Form 12 Version 2, Short Form 12 - Mental Component Score, Patient Activation Measure 13, and Kessler 10

	SF12-PCS (range = 0–100)	*SF12-MCS (range = 0–100)*	*PAM13 (range = 0–100)*	*K10 (range = 10–50)*
Average	48.8	49.1	60.2	17.1
Normality check (sig.)	<0.001	<0.001	<0.001	<0.001
Median	51.8	50.5	55.6	15.0
Minimum	11.4	4.14	0.0	10.0
Maximum	67.7	72.5	100.0	48.0
Quartile 25%	42.7	42.4	51.0	12.0
Quartile 50%	51.8	50.5	55.6	15.0
Quartile 75%	56.2	57.1	67.8	20.0

K10, Kessler 10; MCS, Mental Component Score; PAM13, Patient Activation Measure 13; PCS, Physical Component Score; SF12, Short Form 12 version 2.

Table 3. Validity and Reliability Tests for Short Form 12, Version 2 (Physical Component Score and Mental Component Score), Patient Activation Measure 13, and Kessler 10

	SF12	*PAM13*	*K10*
Construct validity			
Chi-square	-	972.0**	1717.3**
NNFI	-	0.827	0.886
CFI	-	0.867	0.912
RMSEA	-	0.109	0.127
Discriminant validity			
Mean difference between ≤65 and >65 years	–5.5 (PCS) ** –0.8 (MCS)	–0.4	–0.4
Mean difference between non–highly and highly educated	–4.7** (PCS) –1.5**	–7.6**	1.6**
Reliability			
Cronbach α	0.93	0.89	0.92
Guttman λ 6	0.95	0.90	0.93
Split-half	0.90	0.83	0.93

*Sig ≤0.05.
**Sig <0.001.
CFI, Comparative Fit Index; K10, Kessler 10; MCS, Mental Component Score; NNFI, Non-Normed Fit Index; PAM13, Patient Activation Measure 13; PCS, Physical Component Score; RMSEA, Root Mean Squared Error of Approximation; SF12, Short Form 12 version 2.

0.4 and most above 0.6 (Supplementary File S6). Using age as a test for discriminant validity, only the PCS was able to determine a significant difference between the groups younger and older than age 65 (Table 3). The MCS, K10, and PAM13 did not show significant differences between the 2 groups. All tests were able to distinguish between non–highly educated and highly educated individuals. Results were similar in the complete cases and the imputed data.

Reliability

All reliability tests showed similar patterns for all 3 instruments (Table 3). SF12 and the K10 showed a Cronbach α, Guttman λ 6, and Split-Half test of 0.90 or higher, which is good. PAM13 scored lower, but still acceptable. Cronbach α and Guttman λ 6 were 0.89–0.90, and the Split-Half test was 0.83. The results from the complete cases were comparable with the results from the imputed data.

Concepts

An EFA including the 12 health-related characteristics was used to determine the concepts measured (see Supplementary File S5). This analysis provided a KMO test resulting in 0.635, indicating adequate sampling for analysis. KMO tests for individual items were all above 0.5, the acceptable limit, and most above 0.6, and Barlett's test of sphericity was significant ($P < 0.001$). As the scree plot was unclear, eigenvalues less than 1 were used, which led to 4 factors that, when combined, explained 54.5% of variance. Therefore, 4 factors were selected because of the eigenvalues as well as the large sample size. The factor loadings are shown in Table 4. Clustered factors suggest that factor 1 represents mental health, factor 2 represents physical ability, factor 3 represents lifestyle, and factor 4 could stand for self-management. Note that physical exercise is missing in lifestyle and is shown in the fourth factor. In the mental health factor, SF12 MCS and K10 show a strong correlation, indicating they possibly measure the same concept. An additional EFA without K10 still explained 53.2% of variance using the same distribution of the other variables over the same 4 factors. Performing these analyses using only participants younger than age 65 provided similar results. Participants older than age 65, however, gave the same number of factors, but slightly different factor loadings (Supplementary File S5). The influence of employment and disability was omitted, as expected, and health literacy was more related to the mental health factor.

Table 4. Pattern Matrix of Exploratory Factor Analysis of All Health-Related Characteristics

	Mental	*Physical ability*	*Lifestyle*	*Self-management*
Employed	–0.012	**0.385**	–0.031	–0.013
Disabled	–0.163	**–0.374**	0.016	0.049
BMI	0.132	**–0.161**	–0.090	–0.146
Alcohol use	0.057	0.054	**0.628**	0.039
Daily exercise	0.048	–0.043	0.021	**0.227**
Smoking status	0.016	0.060	**–0.350**	0.051
Health literacy	0.085	0.127	0.042	**0.310**
Care use	–0.095	**–0.516**	–0.035	0.005
SF12 PCS	–0.201	**0.856**	–0.041	0.093
SF12 MCS	**0.871**	–0.015	0.020	0.196
PAM13 score	0.001	–0.007	–0.118	**0.644**
K10 score	**–0.722**	–0.240	–0.022	–0.184

Correlations >0.3 are highlighted in *bold*.
BMI, body mass index; K10, Kessler 10; MCS, Mental Component Score; PAM13, Patient Activation Measure 13; PCS, Physical Component Score; SF12, Short Form 12 version 2.

The correlation analyses (Table 5) showed that all measures correlated significantly, but most only had a small overlap. MCS and K10, both mental health measures, showed a strong correlation.

Discussion

This article used a large general population data set to determine the usability of commonly used instruments for

Table 5. Bivariate Pearson Correlation Matrix Between Outcomes

	PCS	*MCS*	*K10*	*PAM13*
PCS	1	0.071**	–0.289**	0.246**
MCS	0.071**	1	–0.779**	0.280**
K10	–0.289**	–0.779**	1	–0.291**
PAM13	0.246**	0.280**	–0.291**	1

*Sig ≤0.05.
**Sig <0.001.
K10, Kessler 10; MCS, Mental Component Score; PAM13, Patient Activation Measure 13; PCS, Physical Component Score.

measuring population health and explored which constructs are measured when they are combined with additional health-related characteristics. Results showed relatively low response rates for SF12 and PAM13, a floor effect in K10, and a deviation in assessed constructs for the PAM13. When the instruments were combined with other health-related characteristics, it became apparent that 4 constructs were measured: mental, physical, lifestyle, and self-management, which matched the intended constructs of the instruments. Despite some shortcomings, which will be discussed further, the results of this study showed that population management initiatives, or other regional and general population-focused policies, can utilize SF12 and PAM13 unaltered in their surveys to evaluate progress on several key aspects of population health.

Data showed that when examining the items with the lowest response rates, their applicability in a general population might be questionable. For example, in a predominantly healthy population, many respondents will answer ''not applicable'' to item 4 (knowledge of medication) of PAM13, as was the case. The high percentages of missing values seen, especially in SF12 and PAM13, are not unique to this study.[33,34] For administrative-based objective measures, such as mortality, this is far less of an issue. However, these are much less able to measure experienced health. Therefore, a solution could be to eliminate questions that are expected to create missing values. For example, the Dutch Health Monitor, a national survey, uses only the first question (general health) of SF12 to measure general health.[35] This question performed better in the present study as well. A more refined alternative would be to tailor surveys. Using item skipping in online surveys, items could be omitted only for those to whom they do not apply. However, this has to be done in consultation with the creators of the instruments and would mean new or additional scoring methodologies. Another option for PAM13, in particular, could be to remove the ''not applicable'' option. The German version of PAM13 merely provides a ''not applicable'' option for item 4 (knowledge of medication), which in a patient population led to better response rates across all other items.[36]

In line with other population-based studies,[37,38] the present study identified that even when instruments' questions were relevant for the general population, most respondents tended to score high when asked about their health. This created positively skewed distributions. This was not problematic for SF12 and PAM13 sum scores as these still provided an acceptable distribution. Conversely, the K10 sum score did show a floor effect, meaning most people had low levels of distress. This lack of distribution in scores makes it difficult for researchers to identify variations between respondents and thus initiatives, making K10 less valuable as an instrument for evaluating PM initiatives. Furthermore, this makes it difficult for insurance companies to potentially use this instrument to award monetary rewards to better performing initiatives. It must be stated that this could be a consequence of the low prevalence of psychological distress in the general population. Additionally, contrary to SF12 PCS and PAM13, which were shown to measure separate constructs, results showed that K10 could be removed from the survey without too much loss of information because of its overlap with SF12 MCS. Some literature even suggests that the SF12 MCS could replace K10 as a screening instrument for depression and anxiety disorders within a general population.[39] These findings imply that 10 questions could be removed from the survey, shortening it and possibly increasing response rates.[15]

Even though this study was able to answer its research questions, some limitations must be acknowledged. First, SF12 uses proprietary scoring software, which ensures correct calculations of SF12 PCS and SF12 MCS. This software, however, prevented an identical analysis strategy for all included instruments as well as the use of imputed data to calculate the sum scores. Therefore, analyses were conducted on both imputed data and complete cases. These did not present remarkable differences throughout the study. The study population itself differed in several aspects from the general Dutch population,[35] which might have affected results. Ideally, even though sensitivity analyses showed it should not have affected conclusions, the influence of these population characteristics should be studied more in depth. All participants answered the instruments in the same order, meaning the effects of survey design on responses could not be studied. Moreover, the EFA conducted suggested the survey did not include several constructs related to the recent broadened concept of health. Qualitative methods, such as end user (citizen) interviews, might have yielded different results, but quantitatively only a few constructs could be distinguished. Predominantly the physical and mental constructs were covered, but not the social construct. In particular, the ability to participate in social activities was lacking.[8] The absence of this construct can be explained by the lack of widely-used validated instruments, but potential instruments are available.[40] Future research should be aware of new definitions of health and consider incorporating new or evaluating existing instruments studying these concepts. This study only assessed the instruments available in the current Dutch setting. Alternatively, a single instrument that encompasses the complete concept of health could be researched. Health is a complex concept and adding more instruments for each of its constructs will lead to unwieldy surveys.

Conclusion

SF12 and PAM13, combined with lifestyle characteristics such as exercise and alcohol use, can be used by PM and other regional initiatives to measure the physical, mental, lifestyle, and health involvement constructs of population health. Furthermore, future population health research should include additional instruments that cover the social construct of health as is defined by new definitions of health.

Acknowledgement

Richard Heijink, PhD provided valuable input for both statistical and textual aspects of this study.

Author Disclosure Statement

The authors declare that there are no conflicts of interest. The authors received the following financial support: This study was funded under SPR project S/133002 of the National Institute of Public Health and the Environment in the Netherlands. The funder had no role in the design of the study, collection, analysis and interpretation of the data and writing of the manuscript.

References

1. Struijs JN, Drewes HW, Heijink R, Baan CA. How to evaluate population management? Transforming the Care Continuum Alliance Population Health Guide into a broadly applicable analytical framework. Health Policy 2015;119:522–529.
2. Berwick DM, Nolan TW, Whittington J. The triple aim: care, health, and cost. Health Aff (Millwood) 2008;27: 759–769.
3. Stiefel M, Nolan K. A guide to measuring the triple aim. Cambridge, MA: Institute for Healthcare Improvement, 2012.
4. Hendrikx RJP, Drewes HW, Spreeuwenberg M, Ruwaard D, Struijs JN, Baan CA. Which triple aim related measures are being used to evaluate population management initiatives? An international comparative analysis. Health Policy 2016;120:471–485.
5. Etches V, Frank J, Di Ruggiero E, Manuel D. Measuring population health: a review of indicators. Ann Rev Public Health 2006;27:29–55.
6. Kindig D, Stoddart G. What is population health? Am J Public Health 2003;93:380–383.
7. World Health Organization. The constitution. Paper presented at International Health Conference, New York, June 19–July 22, 1948.
8. Huber M, Knottnerus JA, Green L, et al. How should we define health? BMJ 2011;343:d4163.
9. Oberhauser C, Chatterji S, Sabariego C, Cieza A. Development of a metric for tracking and comparing population health based on the minimal generic set of domains of functioning and health. Popul Health Metr 2016;14:19.
10. Stevens GA, Singh GM, Lu Y, et al. National, regional, and global trends in adult overweight and obesity prevalences. Popul Health Metr 2012;10:22.
11. Martin L, Nelson E, Rakover J, Chase A. Whole system measures 2.0: a compass for health system leaders. Cambridge, MA: Institute for Healthcare Improvement, 2016.
12. Kimberlin CL, Winterstein AG. Validity and reliability of measurement instruments used in research. Am J Health Syst Pharm 2008;65:2276–2284.
13. Macran S, Weatherly H, Kind P. Measuring population health: a comparison of three generic health status measures. Med Care 2003;41:218–231.
14. van Buuren S. Multiple imputation of discrete and continuous data by fully conditional specification. Stat Methods Med Res 2007;16:219–242.
15. Guo Y, Kopec JA, Cibere J, Li LC, Goldsmith CH. Population survey features and response rates: a randomized experiment. Am J Public Health 2016;106:1422–1426.
16. Drewes HW, Heijink R, Struijs JN, Baan CA. Samen werken aan duurzame zorg. Landelijke monitor proeftuinen. Bilthoven, Netherlands: National Institute for Public Health and the Environment, 2015.
17. Burdine JN, Felix MRJ, Abel AL, Wiltraut CJ, Musselman YJ. The SF-12 as a population health measure: an exploratory examination of potential for application. Health Serv Res 2000;35:885–904.
18. Kessler RC, Andrews G, Colpe LJ, et al. Short screening scales to monitor population prevalences and trends in non-specific psychological distress. Psychol Med 2002;32: 959–976.
19. Rademakers J, Nijman J, van der Hoek L, Heijmans M, Rijken M. Measuring patient activation in The Netherlands: translation and validation of the American short form patient activation measure (PAM13). BMC Public Health 2012;12:577.
20. Ware JE, Kosinski MA, Turner-Bowker DM, Gandek B. User's manual for the SF-12v2® health survey. Lincoln, RI: QualityMetric Incorporated, 2009.
21. Aaronson NK, Muller M, Cohen PD, et al. Translation, validation, and norming of the Dutch language version of the SF-36 health survey in community and chronic disease populations. J Clin Epidemiol 1998;51:1055–1068.
22. Hibbard JH, Mahoney ER, Stockard J, Tusler M. Development and testing of a short form of the patient activation measure. Health Serv Res 2005;40(6 Part 1): 1918–1930.
23. Hibbard JH, Stockard J, Mahoney ER, Tusler M. Development of the patient activation measure (PAM): conceptualizing and measuring activation in patients and consumers. Health Serv Res 2004;39:1005–1026.
24. Tan LSM, Khoo EYM, Tan CS, et al. Sensitivity of three widely used questionnaires for measuring psychological distress among patients with type 2 diabetes mellitus. Qual Life Res 2014;24:153–162.
25. Andrews G, Slade T. Interpreting scores on the Kessler psychological distress scale (K10). Aust N Z J Public Health 2001;25:494–497.
26. Uitewaal P. Depressie in Den Haag. Epidemiologisch Bull 2012;47:23–29.
27. van Buuren S, Groothuis-Oudshoorn K. Mice: multivariate imputation by chained equations in R. J Stat Softw 2011; 45:1–67.
28. Ware JE, Kosinski MA, Keller SD. A 12-item short-form health survey: construction of scales and preliminary tests of reliability and validity. Med Care 1996;34:220–233.
29. CBS; Statistics Netherlands. Monthly labour participation and unemployment. 2017. http://statline.cbs.nl/Statweb/publication/?DM=SLEN&PA=80590ENG&D1=a&D2=0&D3=0&D4=173-187&LA=EN&VW=T Accessed March 8, 2017.
30. CBS; Statistics Netherlands. Bevolking; hoogstbehaald onderwijsniveau en onderwijsrichting. 2017. http://statline.cbs.nl/Statweb/publication/?DM=SLNL&PA=82816ned&D1=0&D2=0&D3=0&D4=0&D5=0-1%2c7%2c11&D6=a&D7=89-92&HDR=G6%2cG3%2cG1%2cG2%2cG4&STB=T%2cG5&VW=T Accessed March 8, 2017.
31. CBS; Statistics Netherlands. Bevolking; kerncijfers. 2017. http://statline.cbs.nl/Statweb/publication/?DM=SLNL&PA=37296NED&D1=a&D2=0,10,20,30,40,50,(l-1)-l&VW=T Accessed March 8, 2017.
32. Swanson RA, Holton EF. Research in organizations: foundations and methods in inquiry: foundations and methods of inquiry. San Francisco: Berrett-Koehler, 2005.
33. Liu H, Hays RD, Adams JL, et al. Imputation of SF-12 health scores for respondents with partially missing data. Health Serv Res 2005;40:905–921.
34. Moljord IE, Lare-Cabrera ML, Perestelo-Pérez L, Rivero-Santana A, Eriksen L, Linaker OM. Psychometric properties of the patient activation Measure-13 among out-patients waiting for mmental health treatment: a validation study in Norway. Patient Educ Couns 2015;98:1410–1417.
35. Brink CLvd, Savelkoul M. Gezondheidsmonitor GGD'en, CBS en RIVM. Volksgezondheid Toekomst Verkenning,

Nationaal Kompas Volksgezondheid. 2013. www.nationaalkompas.nl/algemeen/meta-informatie/bronbeschrijvingen/achtergronddocument-gezondheidsmonitor Accessed January 8, 2016.
36. Zill JM, Dwinger S, Kriston L, Rohenkohl A, Härter M, Dirmaier J. Psychometric evaluation of the German version of the patient activation measure (PAM13). BMC Public Health 2013;13:1027.
37. Andersen LS, Grimsrud A, Myer L, Williams DR, Stein DJ, Seedat S. The psychometric properties of the K10 and K6 scales in screening for mood and anxiety disorders in the South African stress and health study. Int J Methods Psychiatr Res 2011;20:215–223.
38. Bougie E, Arim RG, Kohen DE, Findlay LC. Validation of the 10-item Kessler psychological distress scale (K10) in the 2012 Aboriginal Peoples. 2016. www.statcan.gc.ca/pub/82-003-x/2016001/article/14307-eng.htm Accessed September 6, 2017.
39. Gill SC, Butterworth P, Rodgers B, Mackinnon A. Validity of the mental health component scale of the 12-item short-form health survey (MCS-12) as measure of common mental disorders in the general population. Psychiatr Res 2007;152:63–71.
40. Slabaugh SL, Shah M, Zack M, et al. Leveraging health-related quality of life in population health management: the case for healthy days. Popul Health Manag 2017;20:13–22.

Address correspondence to:
Roy J.P. Hendrikx, MSc
Department for Quality of Care and Health Economics
Center for Nutrition, Prevention and Health Services
National Institute for Public Health and the Environment
PO Box 1
Bilthoven 3720 BA
The Netherlands

E-mail: roy.hendrikx@rivm.nl