Analysis of Questionnaire data using Bayesian Network and Partial Least Squares-Structural Equation Modelling(PLS-SEM)

INTRODUCTION

In this article we show how Bayesian Network(BNs) and Partial Least Squares Structural Equation Model(PLS-SEM) can be developed and used as an effective tool for analysing Questionnaire data .We have conducted a survey from 115 patients to populate our medical dataset.

KEY ATTRIBUTES IN THE DATASET

There are four major constructs: *Care, Burden, Intervention*, and *Adherence*.

Care is a five-item scale consisting of 5 columns (C1 to C5) with values as (poor, fair, good, very good, and excellent).We have done label encoding and mapped (poor,fair,good,very good, and excellent) from 1 to 5.After this we've averaged out the values from C1 to C5 and put it in a column called *Care Average*.

Burden is a 17 item scale consisting of 17 columns from *(BT1: BT17)* with values (A great deal, A lot, Moderate, A small amount, and None). Here we've classified (B1 to B10) as *Primary Burden* and (B11 to B17) is classified as *Secondary Burden*.Similar to what we have done for *Care* we have calculated corresponding *Primary Burden Average* and *Secondary Burden Average*.

Interventions is a 6 item scale consisting of 6 columns (INT1 to INT6) having values (Never, Sometimes, Most of the times, and Always).We've calculated *Intervention average* for each row after performing label encoding.

Adherence is a 5 item "negative worded scale" with 5 columns having values (Never, Sometimes, Most of the times, and Always).Here Always is mapped to 1 and Never is mapped to 5.Similar to Care average, Burden average and Intervention average *Adherence average* is also calculated.

BAYESIAN NETWORKS

Bayesian Networks are also known as recursive graphical models, belief networks,causal probabilistic networks, causal networks and influence diagrams among others(Daly et al. 2011). A BN can be expressed as two components, the first qualitative and the second quantitative (Nadkarni and Shenoy 2001, 2004). The qualitative expression is depicted as a directed acyclic graph (DAG), which consists of a set of variables (denoted by nodes) and relationships between the variables (denoted by arcs) (Salini and Kenett 2009).The quantitative expression comprises probabilities of the variables. Figure 1 shows a Bayesian Network with three variables X, Y and Z. Variables X and Y are parents for variable Z, which indicates that Z is the dependent node. The probability for Z is a conditional probability based on the probabilities of X and Y.

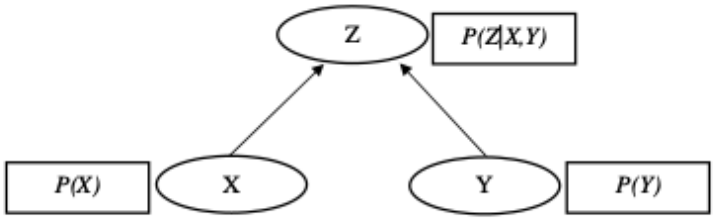


Fig. 1 Example of a Bayesian Network (directed acyclic graph)

MODEL DEVELOPMENT APPROACH

Here we describe our development approach for constructing Bayesian Network models from survey data in four stages: identify the target node, develop the BN structure, formalise the structure of the model, interrogate the model.

Stage I: Identify the target node

In this study we assume that there is one target node which is Adherence.Our main goal is to find what factors can help in increasing the number of cases with high Adherence.

Stage II: Develop the network structure

In order to develop the model we have used GeNIe software that models Bayesian networks,in this software we have chosen PC algorithm to model the BN.

GeNIe Software:

GeNIe Modeler is a development environment for implementing influence diagrams and Bayesian networks, developed at the Decision Systems Laboratory, University of Pittsburgh, and licensed since 2015 to BayesFusion, LLC. Its name and its uncommon capitalization originates from the name Graphical Network Interface, given to the original simple interface to SMILE, a library of classes for graphical probabilistic and decision-theoretic models.GeNIe and SMILE have been originally developed to be major teaching and research tools in academic environments and have been used at hundreds of universities world-wide.Because of their versatility and reliability , GeNIe and SMILE have become very popular and became de-facto standards in academia, while being embraced by a number of government, military, and commercial users.

PC algorithm:

The PC algorithm is a commonly used method for learning the structure of a causal Bayesian network. With a data set, for a pair of variables or nodes (X,Y), the PC algorithm tests their conditional independence given the other variables, and it claims the non-existence of a causal relationship between X and Y, i.e. no edge to be drawn between X and Y, once it finds that X and Y are independent given some other variables. In other words, to determine whether there exists a persistent association between X and Y, the PC algorithm tests the association conditioning on all subsets of all variables other than X and Y. The relationship is considered as causal only when the association exists given each of the conditioning sets.

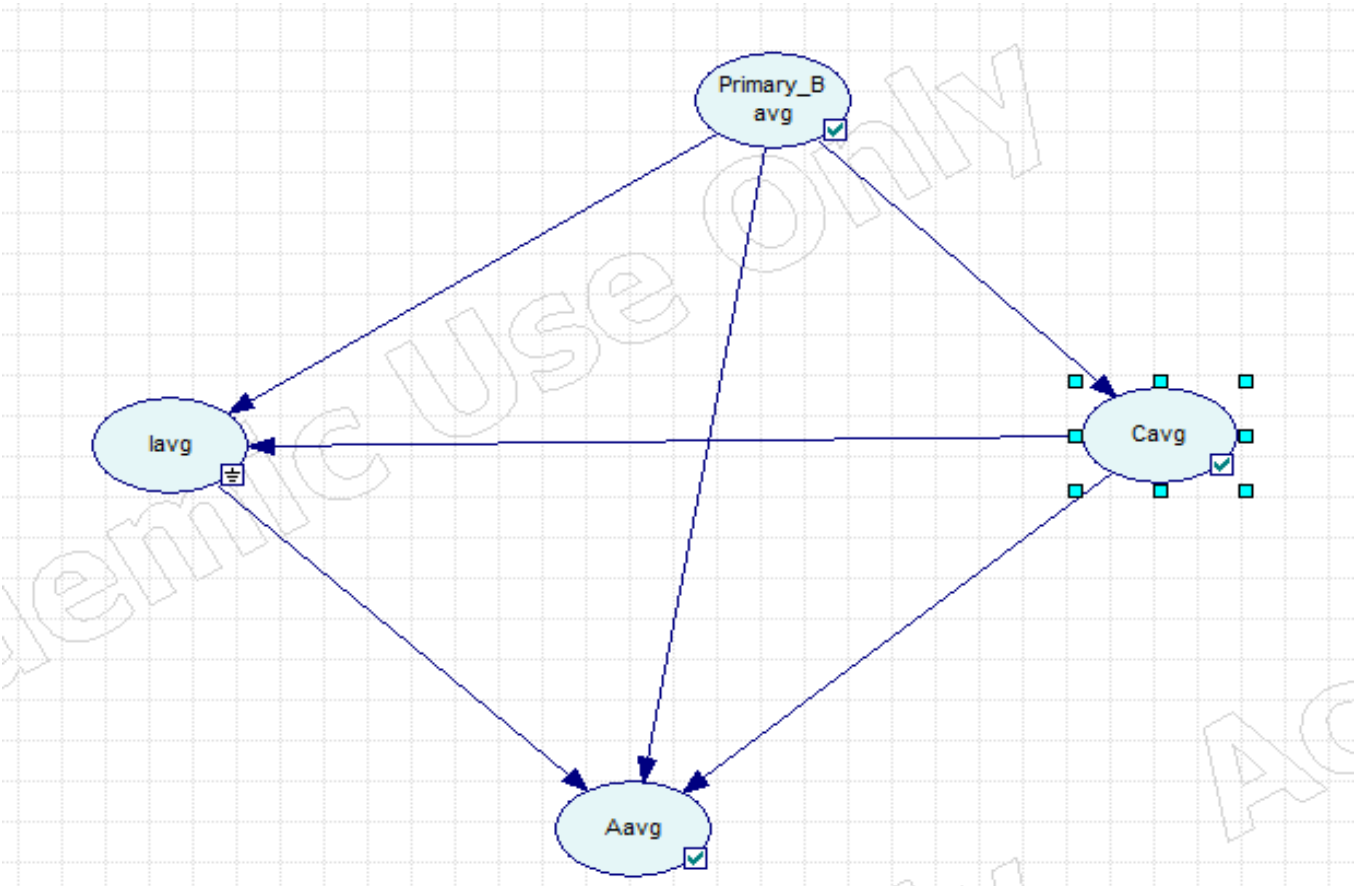
Working of PC algorithm :

Let us consider an exemplar binary data set containing attributes, Gender, College education, High school education, Manager, Clerk, and Salary, where the values of the variables are {male, female} for Gender, {high, low} for Salary, and {yes, no} for all other variables. Note that we use the term target variable to represent the outcome or effect, and predictor variables to represent the inputs or potential causes in this book. For example, variable Salary in this example is the target variable whose value is affected by other variables, and the remaining variables are predictor variables representing possible causes for having a high/low salary. To find out if Gender and Salary have a causal relationship, the independence between Gender and Salary is tested conditioning on each of the subsets of the other variables {College education, High school education, Manager, Clerk}, including the empty set. In the worst case, for each predictor variable, the number of conditional independence tests is 2m−1 where m is the number of predictor variables.

In its algorithmic description, PC starts with a complete (undirected) graph with all the variables, V, and removes the edge between a pair of nodes X and Y immediately after a subset $S \subseteq V \backslash \{X, Y\}$ is found such that X and Y are independent given S. In order to reduce the number of conditional independence tests, the PC algorithm searches for the conditioning set for a pair of nodes in a level by level manner, i.e. searching the conditioning sets with k+1 variables only when the search of all size k conditioning sets fails.

Stage III: Formalise the network structure:

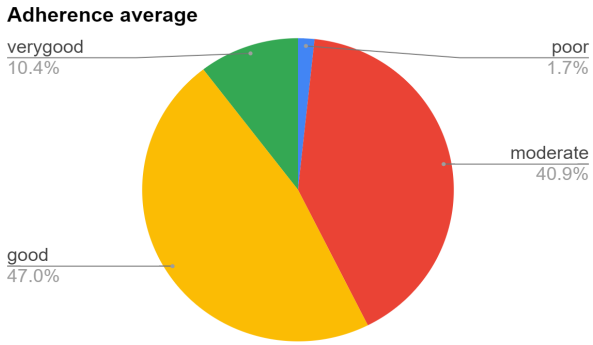
We've identified *Care,Intervention and Burden* as the factors affecting *Adherence*.So in our model we set *Care,Burden* and *Intervention* nodes to point to the Adherence node in order to evaluate the effect of these factors on *Adherence*.The model generated by the software also gave us other arcs between these nodes.

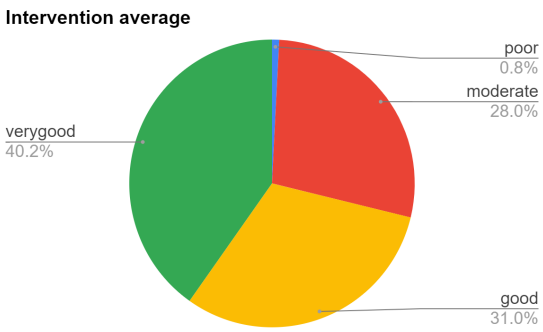


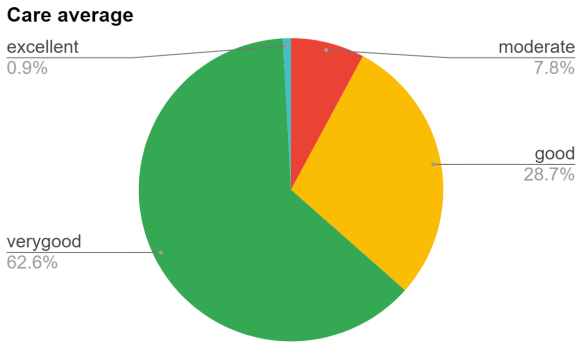
Stage IV: Interrogate the model:

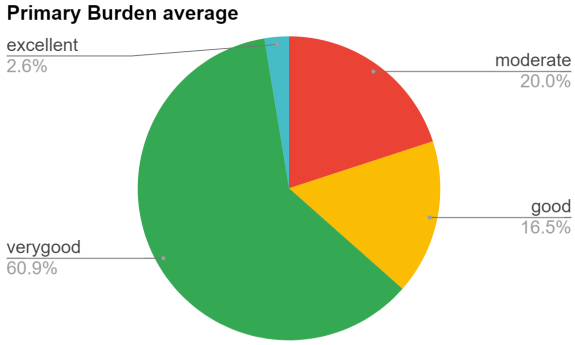
We then tested the complete model by interrogating it.

Current status assessment shows the initial probability settings for all the nodes in the model. After propagating the probabilities through the BN, the top most node *Adherence average* has its poor/moderate/good/very good probabilities equal to **0.017/0.409/0.47/0.104**, and the three key nodes *Intervention average* has poor/moderate/good/very good probabilities **0.008/0.28/0.31/0.402** ,*Care average* and *Primary Burden average* have respective moderate/good/very good/excellent probabilities **0.0783/0.287/0.626/0.0087** and **0.20/0.165/0.609/0.0261** . In the model, the “good” and “very good' state generally has higher probability in all nodes.



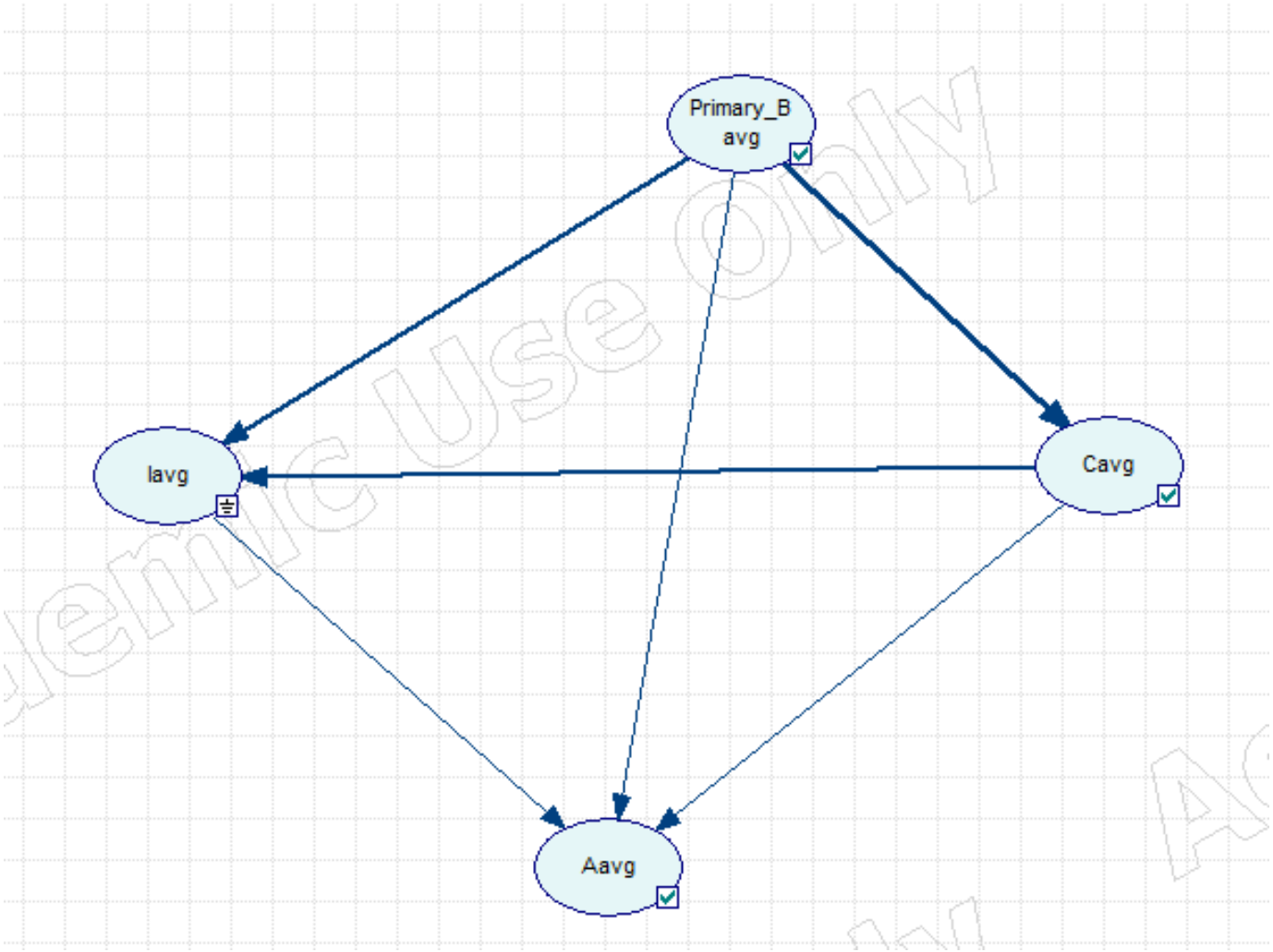






Influence analysis An influence analysis for this model revealed the following:

1. The *Care average* node has a stronger influence on the *Intervention average* node than the *Primary Burden average* node.
2. The *Primary Burden average* node has strong influence on the *Care average* node.



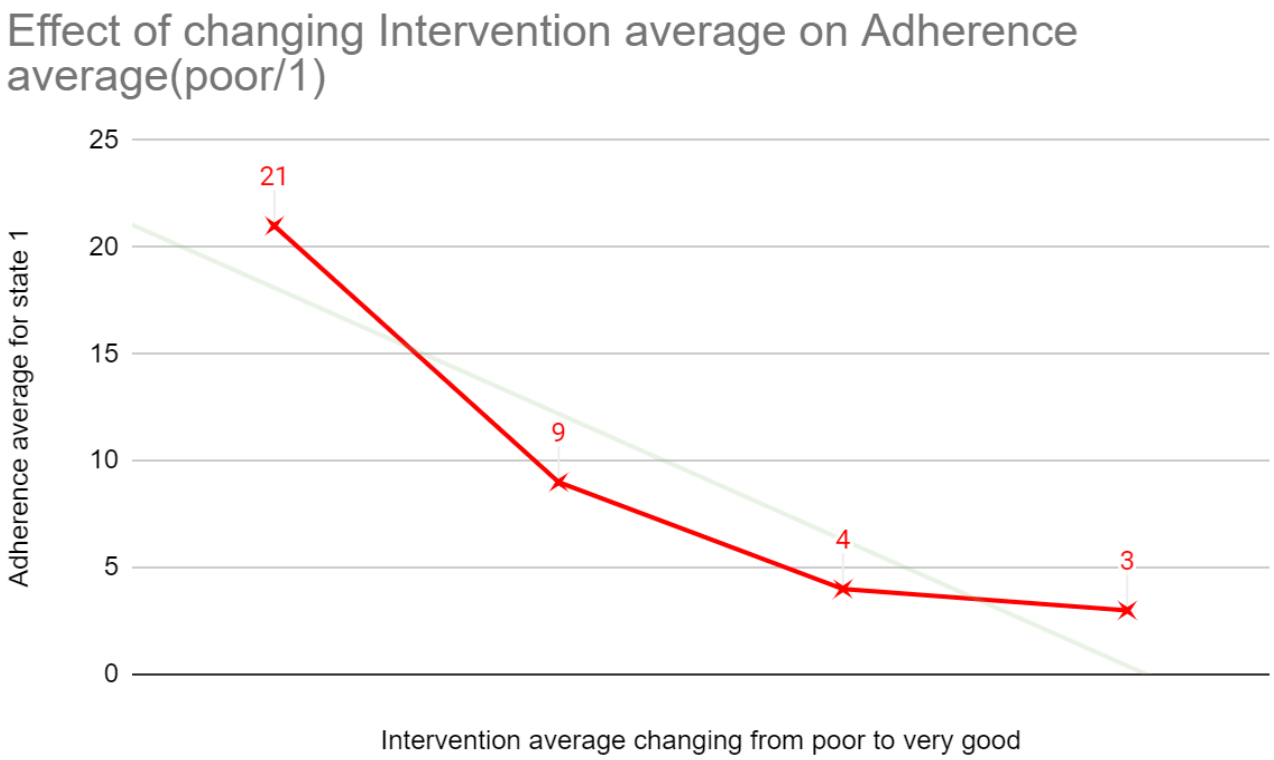
Scenario analyses

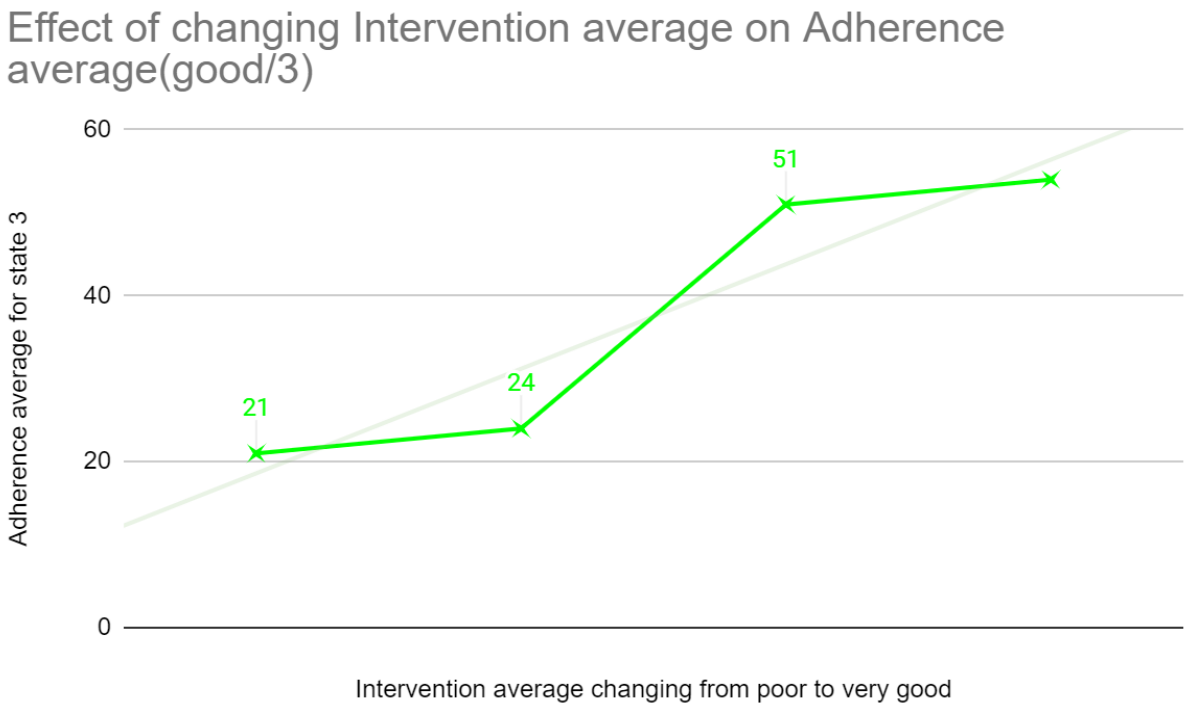
We conducted a series of analyses by changing values of the CPTs according to different scenarios and observing the consequent changes in probabilities in the model. Four such scenarios are presented below.

Scenario 1:

We wanted to see how the number of cases for *Adherence average* states changes as *Intervention average* states varied from poor to very good.We noticed that as Intervention average was set 100% for poor,moderate,good,very good the number of cases with *Adherence average* state as poor(1) decreased from **21% -> 9% ->4%->3%** and the number of cases with *Adherence average* state as good(3) increased from **21% -> 24% ->51%->54% .**

This signifies that as *Intervention average* becomes better (i.e.poor to very good) the number of cases with low states of *Adherence average* decrease and number of cases with high states of *Adherence average* increases.

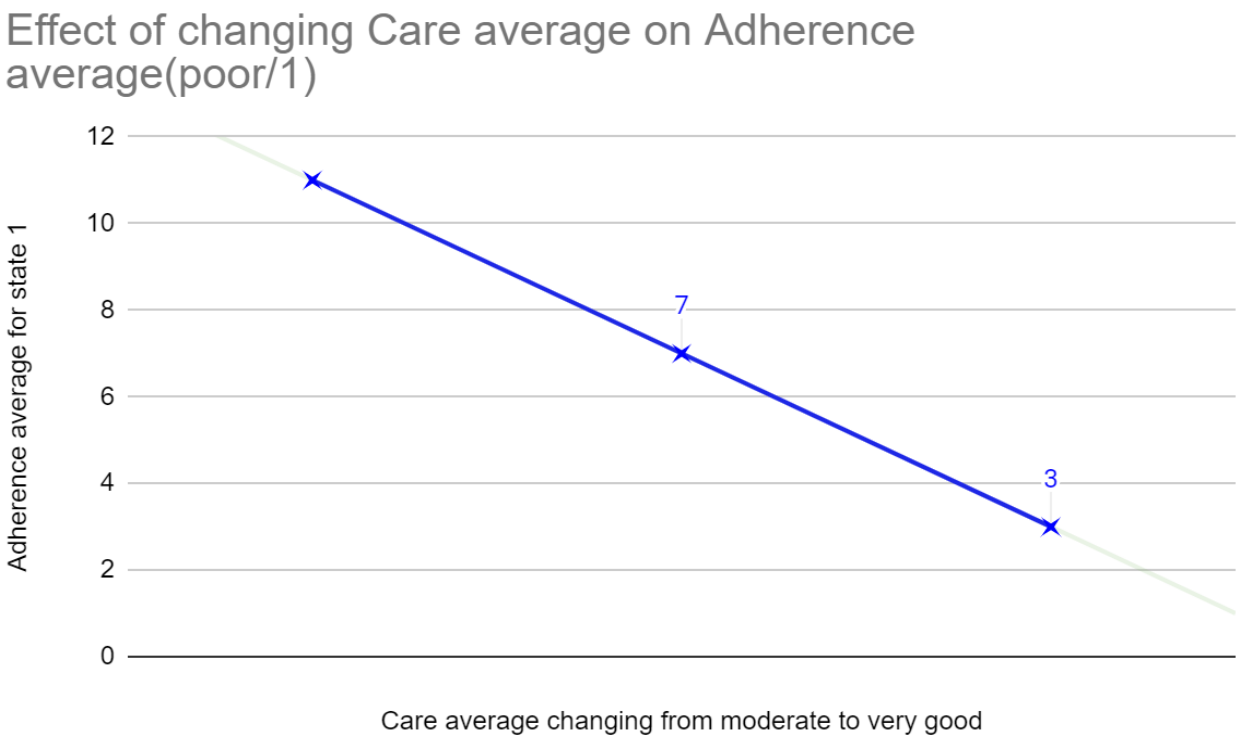


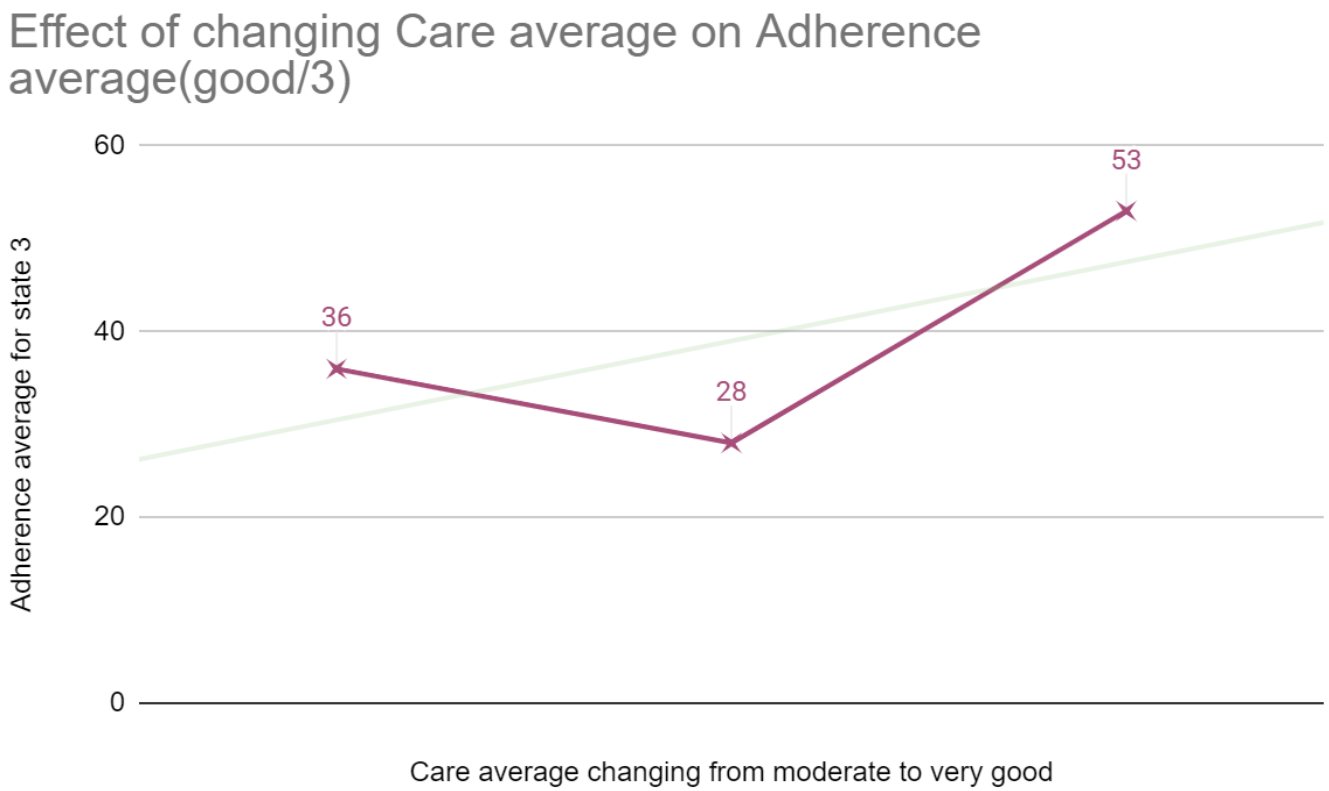


Scenario 2:

We wanted to see how the number of cases for *Adherence average* states changes as *Care average* states varied from poor to very good.We noticed that as *Care average* was set 100% for moderate,good,very good the number of cases with *Adherence average* as state poor(1) decreased from **11% -> 7% ->3%** and the number of cases with *Adherence average* as state good(3) increased from **36% -> 28% ->53%.**

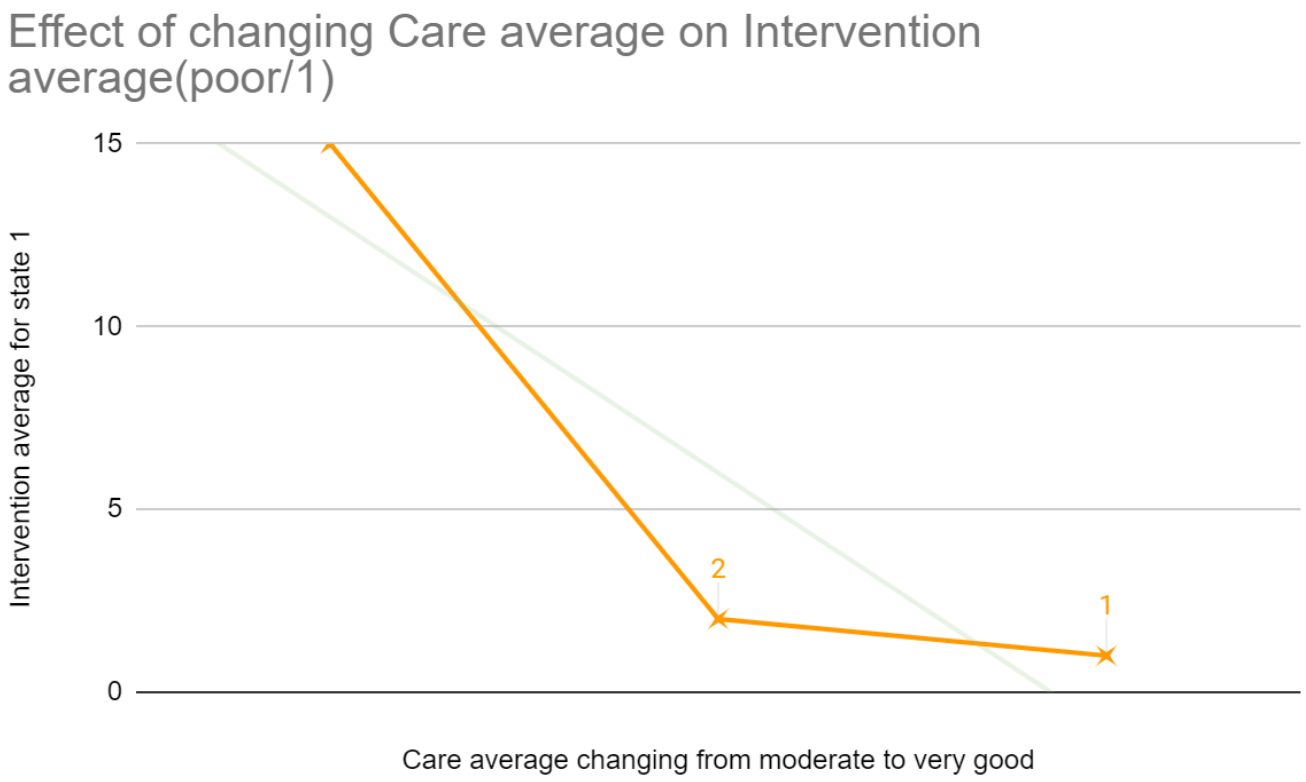
This signifies that as *Care average* becomes better (i.e.moderate to very good) the number of cases with low states of *Adherence average* decrease and high states of *Adherence average* increases.

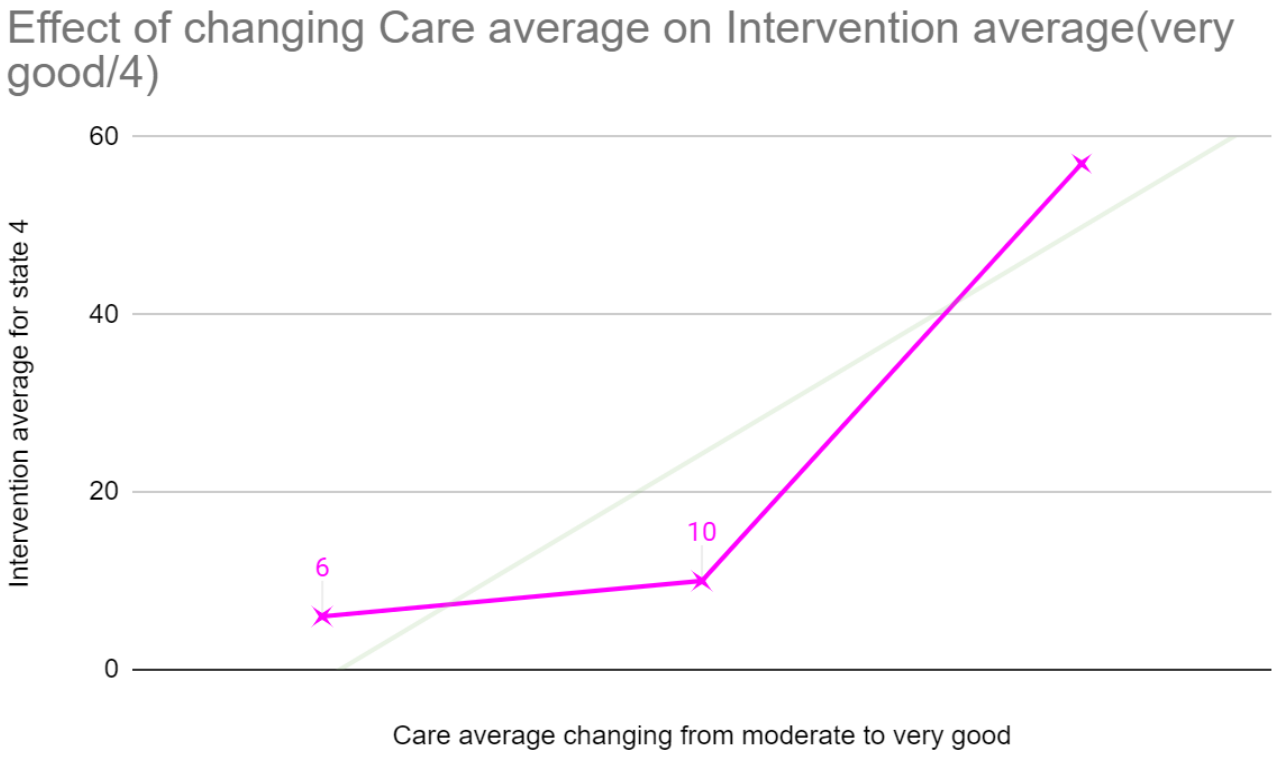

Effect of changing Care average on Adherence average(poor/1)
12
10
8
6
4
2
0
Adherence average for state 1
7
3
Care average changing from moderate to very good


Effect of changing Care average on Adherence average(good/3)
60
40
20
0
Adherence average for state 3
36
28
53
Care average changing from moderate to very good

Scenario 3:

We wanted to see how the number of cases for *Intervention average* states changes as *Care average* states varied from moderate to very good.We noticed that as *Care average* was set 100% for moderate,good,very good the number of cases with *Intervention average* as state poor(1) decreased from **15% -> 2% ->1%** and the number of cases with *Intervention average* as state very good(4) increased from **6% -> 10% ->57%**. This signifies that as *Care average* becomes better (i.e.moderate to very good) the number of cases with low states of *Intervention average* decrease and high states of *Intervention average* increases.

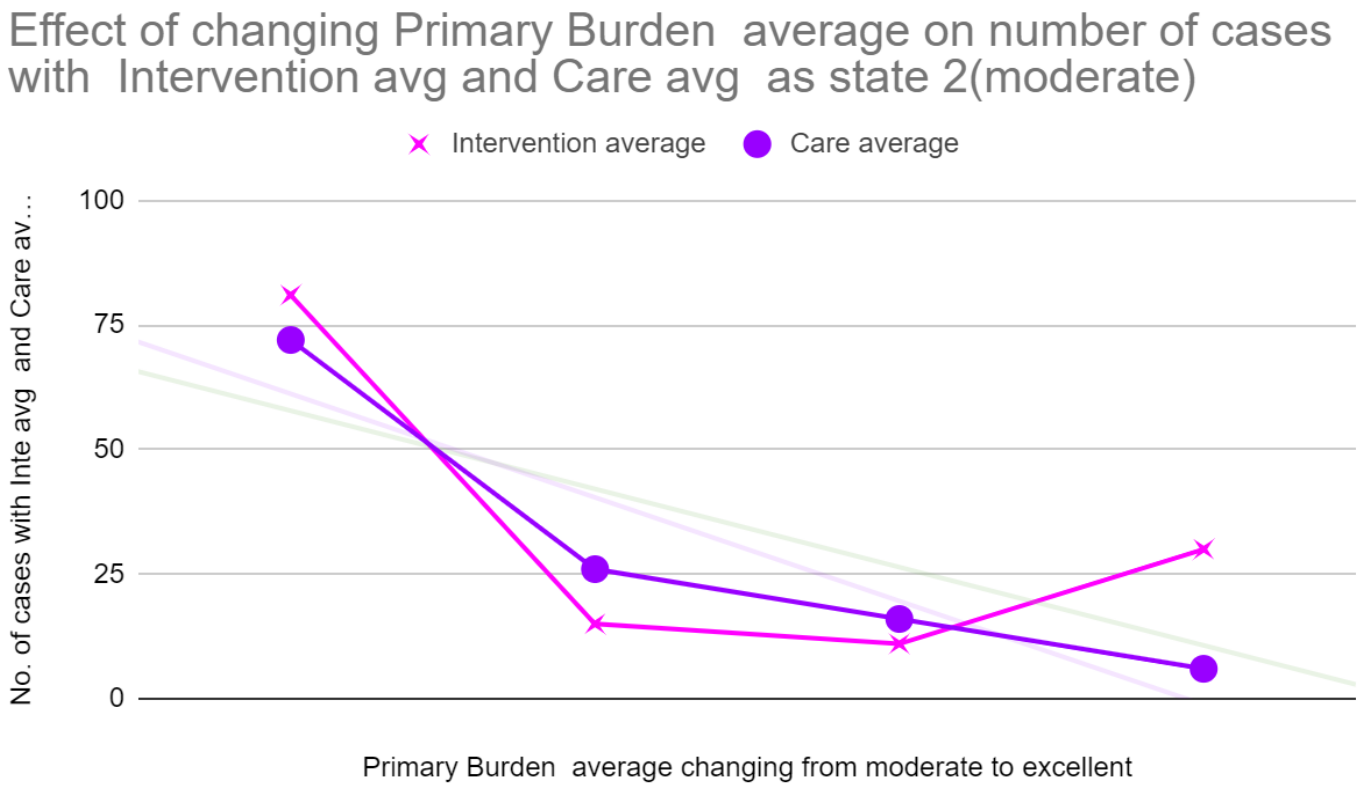


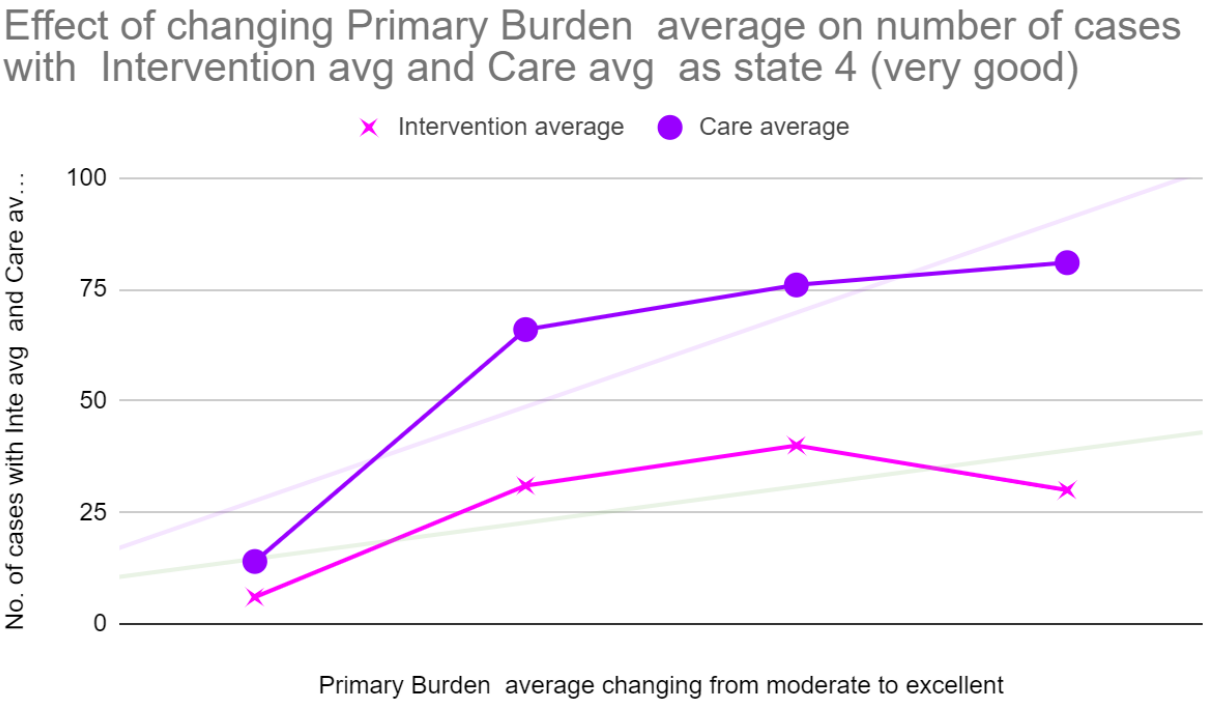


Scenario 4:

We wanted to see how the number of cases of *Intervention average* states and *Care average* states changes as *Primary Burden average* states varied from moderate to excellent. We noticed that as *Primary Burden average* was set 100% for moderate,good,very good, excellent the number of cases having *Intervention average* as moderate(2) decreased from **81% -> 15% ->11%->30%** and the number of cases having *Care average* as moderate(2) decreased from **72% -> 26% ->16%->6%.** The number of cases having *Intervention average* as state 4(very good) increased from **6% -> 31% ->40%->30%** and the number of cases having *Care average* as state very good(4) increased from **14% -> 66% ->76%->81%.**

This signifies that as the Primary Burden average becomes better (i.e.moderate to very good) the number of cases with low states of Intervention average and Care average decrease and number of cases with high states of Intervention average and Care average increases.


Effect of changing Primary Burden average on number of cases with Intervention avg and Care avg as state 2(moderate)
Intervention average
Care average
No. of cases with Inte avg and Care av...
100
75
50
25
0
Primary Burden average changing from moderate to excellent


Effect of changing Primary Burden average on number of cases with Intervention avg and Care avg as state 4 (very good)
Intervention average
Care average
No. of cases with Inte avg and Care av...
100
75
50
25
0
Primary Burden average changing from moderate to excellent

PARTIAL LEAST SQUARES-STRUCTURAL EQUATION MODEL:

PLS-SEM is a causal modeling approach aimed at maximizing the explained variance of the dependent latent constructs. In PLS-SEM approach the correlations between the constructs and their measured or observed variables or items (measuring models) are calculated, and linear regressions between constructs (structural models) are made. PLS-SEM has been increasingly applied in marketing and other business disciplines (e.g., Henseler, Ringle, and Sinkovics 2009) because there are many situations in the applied social and behavioral sciences that are faced with data that do not adhere to a normal multivariate distribution, need more complex models (many constructs and many variables observed), are formative models , have “little” data, and/or are models with less consecrated theoretical support. In these situations partial least square models (PLS-SEM) are recommended (HAIR et al., 2012).

SmartPLS Software

In this paper we will be using the SmartPLS software for PLS-SEM modelling with bootstrapping.SmartPLS is a software with graphical user interface for variance-based structural equation modeling (SEM) using the partial least squares (PLS) path modeling method.Besides estimating path models with latent variables using the PLS-SEM algorithm, the software computes standard results assessment criteria (e.g., for the reflective and formative measurement models, the structural model, and the goodness of fit) and it supports additional statistical analyses (e.g., confirmatory tetrad analysis, importance-performance map analysis, segmentation, multigroup).In bootstrapping, subsamples are created with observations randomly drawn from the original set of data (with replacement).The subsample is then used to estimate the PLS path model.To ensure stability of results, the number of subsamples should be large.

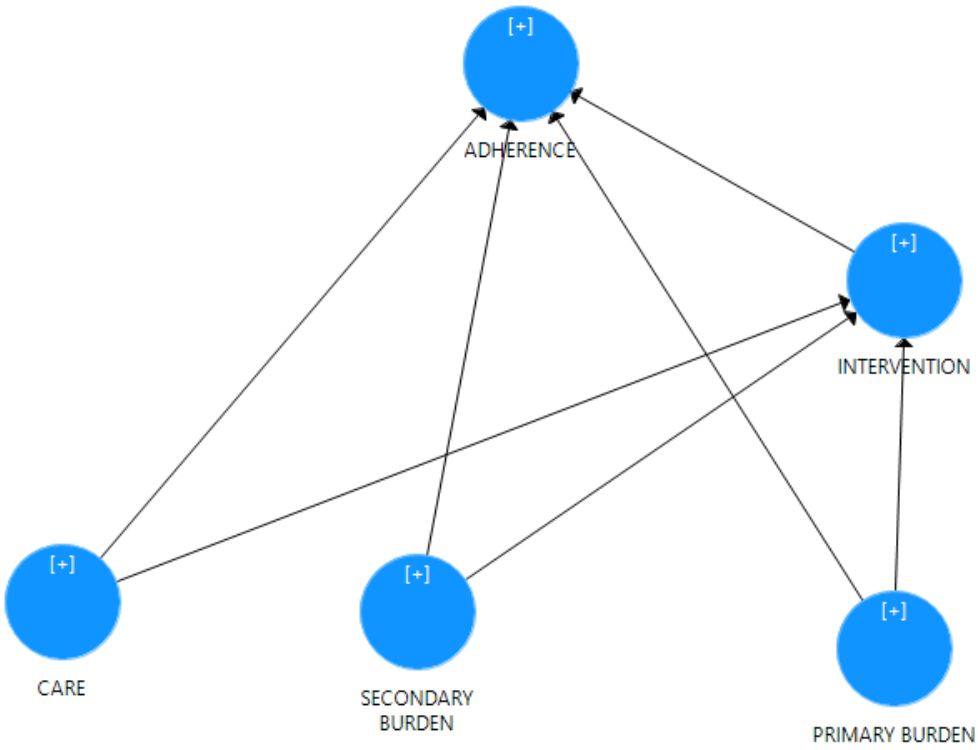
MODEL DEVELOPMENT APPROACH:

Here we describe our development approach for constructing a PLS-SEM model from survey data in two main stages: Model specification that has two sub stages inner model specification and outer model specification,Model evaluation that also has two substages inner model evaluation and outer model evaluation.

Stage I: Model Specification

Inner Model Specification:

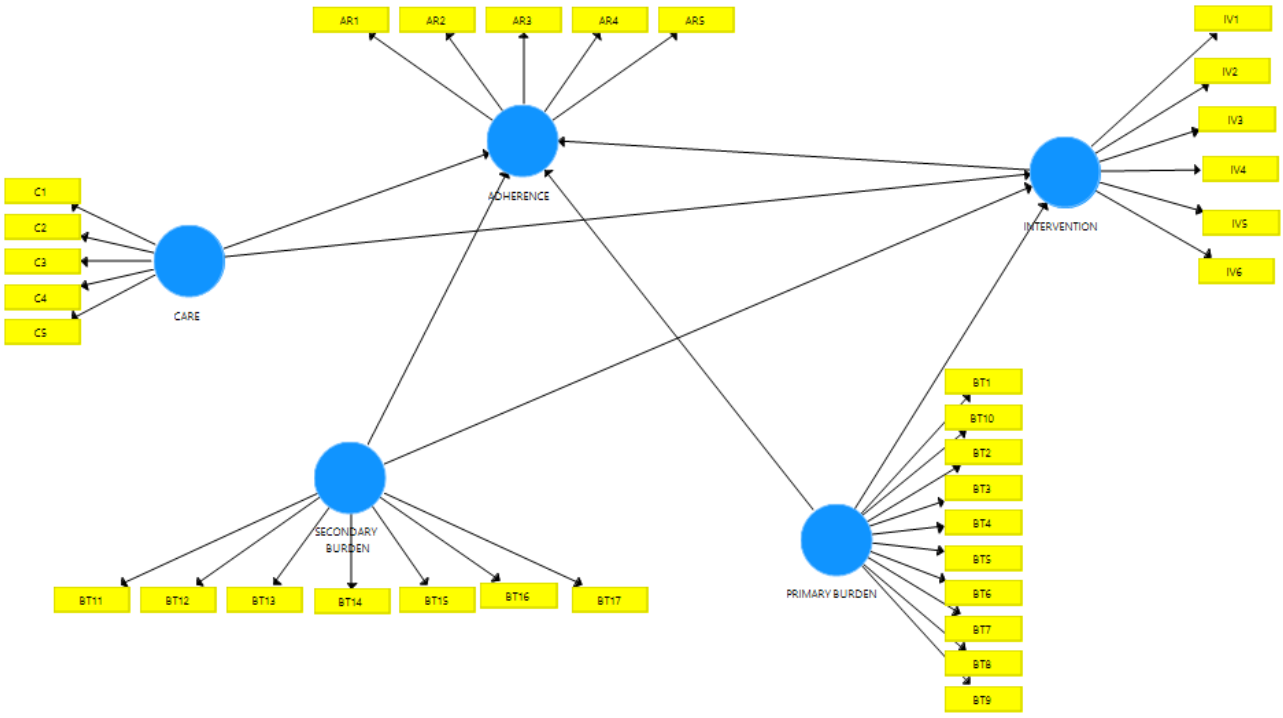
The inner model, or structural model, displays the relationships between the constructs being evaluated.



Here the latent variables Adherence and Intervention are endogenous constructs whereas Care,Secondary Burden and Primary Burden are exogenous constructs.

Outer Model Specification:

The outer models, also known as the measurement models, are used to evaluate the relationships between the indicator variables and their corresponding construct.Our model is a reflective model i.e. the indicators are a representative set of items which all reflect the latent variable they are measuring.



Stage II: Model Evaluation

Conducting evaluation after running PLS-SEM algorithm with bootstrapping for 10000 samples using SmartPLS software,

Outer Model Evaluation:

By starting with the assessment of the outer models, the researcher can trust that the constructs, which form the basis for the assessment of the inner model relationships, are accurately measured and represented. The main criteria to analyse the outer model are based on Hair et al. (2011) as follows:

- Internal Consistency Reliability-Composite reliability should be above 0.6 or 0.7
- Cronbach's alpha should be above 0.6 or 0.7
- Discriminant validity -The AVE of each construct should be greater than the square correlation of the construct with the other latent variables (known as the FornellLarcker criterion)
- The second measure to assess the Discriminant Validity is to look at the cross loadings,In which the outer loading of the construct should be greater than its loadings on other constructs.

LATENT VARIABLE	CRONBACH'S ALPHA	COMPOSITE RELIABILITY
Adherence	0.612	0.760
Intervention	0.832	0.878
Care	0.709	0.811
Primary Burden	0.867	0.893
Secondary Burden	0.840	0.878

The above table demonstrates that the Cronbach's alpha and Composite reliability values are adequate.

We have not considered AVE to prove for convergent validity because we can conclude that convergent validity is adequate based on *Composite reliability* alone because AVE is considered a strict measure of convergent validity. "AVE is a more conservative measure than CR. On the basis of CR alone, the researcher may conclude that the convergent validity of the construct is adequate, even though more than 50% of the variance is due to error." (Malhotra and Dash, 2011, p.702).

The diagonal values represent $\sqrt{AVE\ of\ the\ construct}$.We notice that diagonal values are larger than the respective correlation between the constructs.From the above table we can see that discriminant validity has been established between the given constructs

Cross Loadings	Adherence	Care	Primary Burden	Secondary Burden	Intervention
AR1	0.632	0.245	0.312	-0.236	0.379
AR2	0.605	0.152	0.260	-0.198	0.204
AR3	0.595	0.179	0.233	-0.066	0.239
AR4	0.709	0.204	0.222	-0.166	0.329
AR5	0.570	0.213	0.229	-0.065	0.350
C1	0.292	0.776	0.343	-0.416	0.559
C2	0.262	0.701	0.294	-0.283	0.512
C3	0.160	0.634	0.226	-0.235	0.457

C4	0.201	0.669	0.350	-0.376	0.474
C5	0.180	0.613	0.269	-0.268	0.422
B1	0.267	0.324	0.797	-0.256	0.534
B2	0.292	0.334	0.800	-0.281	0.481
B3	0.173	0.237	0.514	-0.164	0.284
B4	0.331	0.298	0.684	-0.313	0.463
B5	0.316	0.326	0.664	-0.089	0.410
B6	0.174	0.220	0.544	-0.169	0.338
B7	0.345	0.225	0.688	-0.158	0.456
B8	0.387	0.351	0.737	-0.314	0.487
B9	0.113	0.206	0.498	-0.049	0.272
B10	0.247	0.404	0.772	-0.339	0.520
B11	-0.123	-0.218	-0.116	0.676	-0.280
B12	-0.214	-0.245	-0.323	0.748	-0.268
B13	-0.198	-0.483	-0.342	0.809	-0.422
B14	-0.112	-0.276	-0.223	0.730	-0.329
B15	-0.195	-0.388	-0.268	0.762	-0.453
B16	-0.043	-0.217	-0.116	0.460	-0.133
B17	-0.238	-0.417	-0.202	0.775	-0.377
IV1	0.443	0.586	0.509	-0.360	0.796
IV2	0.420	0.579	0.526	-0.373	0.815
IV3	0.283	0.490	0.463	-0.414	0.726
IV4	0.400	0.444	0.497	-0.384	0.721
IV5	0.309	0.590	0.510	-0.324	0.739
IV6	0.356	0.471	0.315	-0.269	0.625

The above table satisfies discriminant validity conditions.The outer loading of the construct is greater than its loadings on other constructs.

Inner Model Evaluation:

The main criteria to analyse the outer model are based on Hair et al. (2011) as follows:

- Path coefficients between latent constructs (p value should be <0.05)
- R^2 value should be >0.15 for medium effect and >0.35 for strong effect in social science research

In order to assess path coefficients we look at direct and indirect effects between the latent constructs.

Direct Effects	P value
Care -> Adherence	0.526
Care -> Intervention	0.000
Intervention -> Adherence	0.003
Primary Burden ->Adherence	0.188
Primary Burden ->Intervention	0.000
Secondary Burden ->Adherence	0.858
Secondary Burden ->Intervention	0.140

From the above table we can infer that

- *Care* influences *Intervention*
- *Intervention* influences *Adherence*
- *Primary Burden* influences *Intervention*.

Indirect Effects	P-value
Care->Intervention->Adherence	0.008
Primary Burden->Intervention->Adherence	0.017
Secondary Burden->Intervention->Adherence	0.219

From the above table we can infer that

- *Intervention* mediates the effect of *Care* on *Adherence*
- *Intervention* mediates the effect of *Primary Burden* on *Adherence*
-

R square	Original Sample(O)
Adherence	0.267
Intervention	0.676

The table demonstrates that the values are adequate with Adherence and Intervention indicating medium and strong effects respectively.

Stage III: Multi Group Analysis

Through this analysis we will try to find out how social factors affect relationships between the latent constructs in our model.

- Under the Funding type column we have two groups-Intervened and Self
- For multi group analysis we will individually look at :
  - Path coefficients of the latent constructs
  - R^2 value of endogenous constructs
  - Latent variable correlations

Category 1 : Funding Type-Intervened

Direct effects	P Value
Care->Intervention	**0.000**

Latent Construct	R Squared	Adjusted R Squared
Adherence	0.249	0.190
Intervention	0.492	0.463

Latent Variable Correlations	Original Sample(O)	Sample Mean(M)	STDEV	T-Statistics	P Value
Intervention->Care	0.697	0.712	0.094	7.455	**0.000**
Secondary Burden-> Care	-0.612	-0.533	0.214	2.862	**0.004**

Category 2: **Funding Type-Self**

Direct and Indirect effects	P Value
Care->Intervention	**0.000**
Intervention->Adherence	**0.002**
Primary Burden->Intervention	**0.000**
Secondary Burden->Intervention	**0.022**
Care->Intervention->Adherence	**0.024**
Primary Burden->Intervention->Adherence	**0.003**

Latent Construct	R Squared	Adjusted R Squared
Adherence	0.366	0.319
Intervention	0.753	0.740

Latent Variable Correlations	Original Sample(O)	Sample Mean(M)	STDEV	T-Statistics	P Value
Care->Adherence	0.363	0.386	0.107	3.395	**0.001**
Intervention->Adherence	0.601	0.614	0.072	8.334	**0.000**
Intervention->Care	0.654	0.674	0.065	9.988	**0.000**

Primary Burden->Adherence	0.460	0.476	0.100	4.613	**0.000**
Primary Burden->Care	0.447	0.474	0.107	4.170	**0.000**
Primary Burden->Intervention	0.787	0.796	0.044	18.034	**0.000**
Secondary Burden->Intervention	-0.478	-0.506	0.108	4.407	**0.000**
Secondary Burden->Adherence	-0.427	-0.451	0.099	4.296	**0.000**

From the above multigroup analyses we can infer that for the category Self:

- ❏ The R^2 values are greater than for Intervened Category
- ❏ There are more path coefficients with p value less than 0.05 when compared to Intervened Category.

In the Self category we can notice that there is an indirect effect between Primary Burden and Adherence through the latent construct Intervention.But this same indirect effect is not present in the Intervened Category.Therefore we can infer that the Primary Burden and the Funding category moderate the effect of Interventions on Adherence.

HYPOTHESES :

- Hypothesis 1 (H1): Primary Caregiver Perceived Value (PCPV) of physician influences the design, choice, and administration of individualized interventions
- Hypothesis 2 : Individualized interventions mediate the effect of PCPV of the physician on the primary caregiver's adherence to their child's treatment.
- Hypothesis 3: The contextual factors that comprises of primary caregiver's treatment burden moderate the effect of PCPV of the physician on the selection and administration of individualized interventions.
- Hypothesis 4 : The contextual factors (primary caregiver's treatment burden) and the social factors (that comprises primary caregiver's family resources (social, cultural, religious, economical, educational and medical resources)) moderate the effect of individualized interventions on Adherence to treatment.

HYPOTHESIS	BAYESIAN NETWORKS	PLS-SEM
H1	**Proven in** *Scenario 1*,*Scenario 2*,*Scenario 3*	**Proven in** Inner Model Evaluation
H2	**Proven in** *Scenario 3*	**Proven in** Inner Model Evaluation
H3	**Proven in** *Scenario 4*	
H4		**Proven in** Multi Group Analysis

REFERENCES:

We have adapted the steps for the Bayesian Network from the paper (Chakraborty 2016).(https://link.springer.com/article/10.1186/s40165-016-0021-2) The paper formalises and presents an innovative general approach for developing complex system models from survey data by applying Bayesian Networks. The challenges and approaches to converting survey data into usable probability forms are explained and a general approach for integrating expert knowledge (judgements) into Bayesian complex system models is presented. The structural complexities of the Bayesian complex system modelling process, based on various decision contexts, are also explained along with a solution. A novel application of Bayesian complex system models as a management tool for decision making is demonstrated using a railway transport case study. Customer satisfaction, which is a Key Performance Indicator in public transport management, is modelled using data from customer surveys conducted by Queensland Rail, Australia.

The steps for the PLS-SEM model have been adapted from the PLS-SEM:Indeed silver Bullet Paper(http://scholar.google.co.in/scholar_url?url=http://www.academia.edu/download/42526295/Hair_et_al_2011.pdf&hl=en&sa=X&scisig=AAGBfm2c_1A5IsQ71Nz-U1ng9CuxQNHk3w&nossl=1&oi=scholarr).The paper reviews PLS-SEM and its algorithm, and provides an overview of when it can be most appropriately applied, indicating its potential and limitations for future research. The

authors conclude that PLS-SEM path modeling, if appropriately applied, is indeed a “silver bullet” for estimating causal models in many theoretical models and empirical data situations.