The Role of Health Status and Social Capital in Cancer Mortality: Insights from Matched Register and Survey Data

ELINE AAS[1,2]
CHRISTIAN B. H. THORJUSSEN[1,3,5]
GEIR GODAGER[1,4]

[1] Department of Health Management and Health Economics, Institute of Health and Society, University of Oslo, Oslo, Norway
[2] Division for Health Services, Norwegian Institute of Public Health, Oslo, Norway
[3] Nofima AS, Ås, Norway
[4] Health Services Research Unit, Akershus University Hospital, Oslo, Norway
[5] Norwegian University of Life Sciences, Ås, Norway

Abstract: Cancer mortality has been shown to be associated with social- and human capital. Several channels have been suggested, such as early detection, better compliance to treatment and better health prior to diagnosis. In this paper we study how health status and social capital jointly affect cancer mortality and cancer severity at the time of diagnosis. The analyses are based on study sample of individuals with cancer diagnosis. Our merged dataset contains information on cancer diagnosis and death from the Cancer Registry of Norway and health status, social capital and other individual level data from several national health surveys measured before the time of diagnosis. Health status and social capital are treated as unobserved latent variables, and we apply generalized structural equation modelling framework to estimate conditional statistical associations of social capital and individual health on cancer severity and mortality. We find that health has negative, and statistically significant effect, on cancer mortality, while we cannot conclude on the association between health and cancer severity (metastasis yes/no). We cannot conclude that cancer mortality and the probability of cancer metastasis are associated nor disassociated with social capital. Our results add nuance to prior studies, which frequently report a significant association between social capital and cancer mortality.

JEL classification: C31, C50, H44, I10, I14

Keywords: cancer mortality, health status, social capital, general structural equation modelling

1. Introduction

Worldwide, cancer is one of the most recurrent diseases, and left untreated, the consequence of cancer is often fatal. We revisit the question of how cancer mortality relates to individuals' general health status and social capital. Severity at the time of diagnosis is closely related to survival. Five-years relative survival in distal (the most severe group), colon (ICD-10 C18), and lung (trachea ICD-10 C33-34) cancer are about 17 percent and 5 percent, respectively, while in local cancers (least severe stage) it is 97 percent and 67 percent, respectively (Cancer Registry

**** Correspondence to:*** Eline Aas, Department of Health Management and Health Economics, University of Oslo, Oslo, Norway. E-mail: eline.aas@medisin.uio.no.

Published: Online December 2023. dx.doi.org/10.5617/njhe.10304

of Norway, 2021). With a strong relation between survival and severity, it is important that individuals are diagnosed as early as possible.

Several factors have been shown to influence access to healthcare, measured by severity at the time of diagnosis, and survival, such as social capital and health status. Social capital is a concept with several dimensions (Islam, 2006, Putnam, 2004 and Folland, 2014). First, social capital is divided into cognitive and structural components. While cognitive social capital refers to norms, values, attitudes and beliefs, structural social capital contains observable aspects of social organization, such as membership in formal voluntary organizations and participation in informal networks. Structural social capital and cognitive social capital are likely to interact. Second, in empirical studies there is often a distinction between individual and community social capital. This distinction corresponds to the distinction between variables at the individual level and the community level. While community social capital (CSC; for instance, total number of memberships in voluntary organizations) informs us about the aggregate level of interactions and networks in the community irrespective of whether a particular individual is part of the network, individual social capital (ISC; measured for instance by marital status, the number of networks and formal organizations in which an individual is a member) indicates the social capital of this particular person.

The effect of social capital has been included in several studies. Two population-based studies from Norway of 12 common types of cancer (Kravdal, (2000) and (2001)) have identified social differentials in survival. Kravdal found that all-cause mortality among cancer patients compared with similar patients without a cancer diagnosis were lower among cancer patients married, with higher education, having an occupation and with high income. This protective effect of marriage was not due to stage, which was controlled for. Similar effects of marital status have been shown in Goodwin et al. (1987), Johansen et al. (1996), Villingshøy et al (2006), Wang et al (2011), Auvinen et al. (1995) and Aizer et al (2013).

Kravdal discusses possible mechanisms that may contribute to the explanation of the association. He suggests that a spouse and children may take the initiative to obtain a second opinion about the diagnosis, which may be important for the treatment. Second, they may get involved in the choice of type of treatment. Third, the type of treatment that is chosen may be rationed. Pressure from the family may have an influence on the priority a patient obtains. Fourth, patients with family may be helped to follow instructions more accurately, and to take initiative for further consultation if they notice signs of recurrence or other problems during periods when patients are under less close medical surveillance. Finally, Kravdal (2001) suggests that the married patients may possibly be able to attract more treatment resources, and more formal care resources. Similar mechanisms have been discussed by Goodwin et al (1987), who suggest that the favorable consequence of being married on overall survival is due to multiple beneficial effects; early diagnosis, choice of treatment, and response to treatment all seem to have effect.

Better health status prior to a cancer diagnosis could result in improved cancer survival due to less complications during treatment, which would improve compliance to treatment guidelines. Compliance to treatment guidelines would be expected to improve survival. Examples could be recovery after surgery and less complications during chemotherapy. Health status is correlated with socioeconomic status, hence the effect of socioeconomic status on access to healthcare and survival, might be influenced when adjusting for health status.

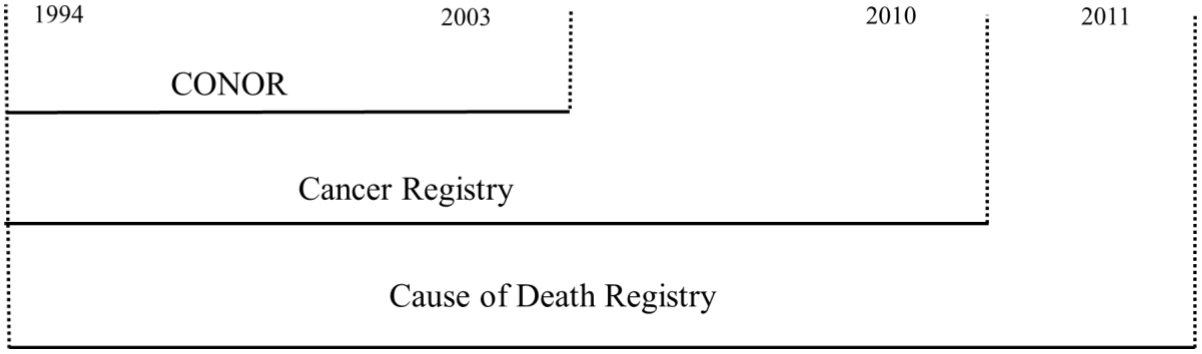
In most healthcare systems, equal access to the healthcare services is important. Hence, the objective of this paper is to study equity in the provision of cancer treatment, by addressing equality in access to care by 5-year mortality. We suggest that a higher degree of social capital and better health status reduce 5-year mortality rate, after simultaneously controlling for metastasis at the time of diagnosis. Although the results from the existing literature vary somewhat; the trend seems to be that social capital has a positive impact on cancer survival.

The present study adds to the literature by adjusting for health status and employing data that makes it possible to estimate the contribution to survival from a wide range of variables. In addition, we model social capital and health as latent constructs, in a generalized structural equation model (GSEM).

2. Data

We base our analysis on the national database Cohort of Norway (CONOR). The CONOR database include data from 10 different Norwegian epidemiological studies. Collection of data started in 1994 in Tromsø, and continued in the counties Nord-Trøndelag (1995-97), Hordaland (1997-99), Oslo (2000-01), Oppland and Hedmark (2000-01), Tromsø (2001), Oslo (2002), Troms and Finnmark (2002) and Oslo (2003). By December 2003, the data included information on 173,236 individuals. The CONOR studies collected a whole array of information including answers to 50 questions related to family and social life, home environment, work, health, physical activities, and other factors expected to influence health and wellbeing. The frame of the questions varies slightly between the epidemiological studies. Answers are recoded to ensure correspondence between studies. The CONOR data are sampled from an adult Norwegian population. Each study targeted everyone in a county within different age groups. Since participation was voluntary, the possibility of systematic selection bias driven by observables was investigated. We do not suspect substantial sample selection biases, as analysis shows that the CONOR data is representative of the Norwegian population within the same age groups surveyed in CONOR (Aamodt, et al. 2015). Furthermore, we have compared several measures (such as life-expectancy, proportion with higher education, average household income, and proportion of smoking in women) for the total Norwegian population with counties that are represented in our study sample. One may argue that the study sample is a fair representation of the Norwegian population (see Appendix Table A1).

Figure 1: Timeline for all data sources.



The CONOR data was merged with data from the Cancer Registry of Norway (NCR) to select the study sample. From NCR we collected information on time of diagnosis, type of cancer and metastasis. To avoid problems with simultaneity bias caused by endogeneity, we used information from the CONOR survey *prior* to the cancer diagnosis. The final study sample consisted of individuals with a cancer diagnosis and with information from the CONOR survey collected either the same year as the cancer diagnosis or maximum 7 years prior to the cancer diagnosis. For example, in our data sample there are individuals with a cancer diagnosis in 2000 with information from the CONOR survey from 1999 and individuals with a cancer diagnosis in 2007 with information from the CONOR survey from 2003 (four years prior to diagnosis). In addition, we merged information from the Norwegian Cause of Death Registry (NCDR) with details on the cause (ICD-10) and time of death. Our sample data included 9179 individuals with a known cancer diagnosis and with information from the CONOR survey. When estimating 5-year mortality, we include all causes of death. Figure 1 describes the time coverage of the data: CONOR questionnaires were sampled from 1994 to 2003, cancer data from 1994 to 2011, and mortality data from 1994 to 2011.

An important feature of our data, which distinguishes our contribution from previous literature, is that the CONOR questionnaire answers were collected before any of the individuals knew they had cancer.

Table 1: Definition of variables.

Variable	Description
Metastasis	The primary cancer for a given individual is metastatic.
5-year mortality	The person died of any cause within five years after primary cancer diagnosis.
Second cancer	A second cancer diagnosed within five years after primary cancer.
Smoking	The person is a daily smoker at the time of the CONOR questionnaire, zero otherwise.
Health status	Self-assessed health encoded as 1 = very poor, 2 = poor, 3 = good, 4 = very good.
Male	The person is a man
Age	Age of person at diagnosis
Higher Education	The person has finished education from college or university
Marital status:	Self-reported marital status
Married	Person is married or lives with partner
Unmarried	The person has never married
Widow	The person is a widow
Divorced	The person is divorced
Separated	The person is separated
Enough friends	The person feels that he or she has enough friends.
Number of friends	Self-reported number of friends encoded as follows, 0 = (0-4), 1 = (5-9), 2 = (10-14), 3 = (15 – 19), 4 = (20 – 24), 5 = more than 25.
Employed	The person is permanently employed, a student or in the military service.
Activities	The person participates in organized activities at least once a week.
No heart Attack	The person has never suffered a heart attack prior to the cancer diagnosis.
Light physical activity	Self-reported answer to following question; how often do you do physical activity per week, where you were not sweating or out of breath?
Hard physical activity	Self-reported answer to following question; how often do you do physical activity per week, where you were sweating or out of breath?
No stroke	The person has never suffered a stroke prior to the cancer diagnosis.
No diabetes	The person never ever had diabetes (either type 1 or type 2)
Lung cancer	The primary cancer is lung cancer. ICD-10 codes: C33-34
Digestive cancer	The primary cancer is in the digestive system. ICD-10 codes: C15-17, C22-26
Unspecific cancer	The primary cancer is ambiguous usually due to several tumors. ICD-10 codes: C39, C76, C80
Skin cancer	The primary cancer is skin cancer. ICD-codes: C43-44
Breast cancer	The primary cancer is breast cancer. ICD-codes: C50
Genital	The primary cancer is in the female genitals or ovary. ICD-codes: C51-58
Prostate cancer	The primary cancer is prostate cancer. ICD-codes: C61
Urinary cancer	The primary cancer is in the urinary system. ICD-codes: C64-68
Central nervous system	The primary cancer is in the central nervous system. ICD-codes: C70-72, D42-43
Lymph cancer	The primary cancer is in the lymphatic system. ICD-codes: C81-85, C88, C90-96, D45-47
Colon cancer	The primary cancer is in the colon. ICD-codes: C18-21
Other	The primary cancer is not any of the above. ICD-codes: C00-14, C30-32, C37, C74-75, C38, C40-41, C45-46, C48-49, C60, C62, C63, C69, C73
Diagnostic month	Month in which cancer was diagnosed. Dummy variable for each month where January is the reference month

This feature of the data enables us to measure health and social capital pre-dating any known cancer diagnosis, thereby identifying the effect of health and social capital on metastasis and mortality. In Table 1, we give a description of the variables included in our model, in Appendix A.1 we provide some summary statistics for these variables.

2.1. Descriptive statistics

In Table 2 we have reported the sample in total and subsamples of individuals with metastasis and dead.

Table 2: Descriptive statistics describing sample size mean and standard deviation min and max of variables in Table 1.

		Total sample			Metastasis			Dead		
Variable	**Category**	**#**	**Mean**	**St.dev**	**#**	**Mean**	**St.dev**	**#**	**Mean**	**St.dev**
Age		9179	67.5	12.8	2568	68.8	11.9	4046	71.9	11.1
Gender	Male	9179	0.565	0.496	2568	0.571	0.495	4046	0.615	0.487
Employed	Yes	9179	0.324	0.468	2568	0.282	0.450	4046	0.198	0.399
Education	Higher education	8631	0.189	0.392	2408	0.156	0.363	3736	0.130	0.337
Activities	No	9179	0.130	0.336	2568	0.129	0.335	4046	0.112	0.315
Friends	No	6427	1.333	1.312	1762	1.299	1.295	2656	1.304	1.318
Lonely	Not lonely	6657	2.719	0.604	1899	2.700	0.629	2886	2.664	0.661
Marital status	Married	9117	0.661	0.473	2553	0.655	0.475	4026	0.636	0.481
	Unmarried	9117	0.101	0.301	2553	0.095	0.293	4026	0.092	0.290
	Widow	9117	0.140	0.347	2553	0.159	0.366	4026	0.186	0.389
	Divorced	9117	0.087	0.282	2553	0.080	0.271	4026	0.077	0.266
	Separated	9117	0.012	0.107	2553	0.011	0.106	4026	0.009	0.095
Comorbidity	No stroke	8974	0.956	0.204	2553	0.957	0.204	3948	0.941	0.235
	No diabetes	9007	0.944	0.230	2519	0.945	0.229	3971	0.926	0.262
	No heart attack	9002	0.924	0.265	2524	0.914	0.280	3973	0.897	0.304
Physical activity	Light	8082	3.114	0.991	2272	3.062	1.034	3489	3.031	1.057
	Hard	7177	1.915	1.043	1985	1.829	1.028	3035	1.742	1.005
Health status	Scale 1-4	9067	2.688	0.678	2538	2.636	0.671	3983	2.588	0.674
Smoking	Daily	9059	0.298	0.457	2536	0.368	0.482	3981	0.345	0.475
Metastasis	Yes	9179	0.280	0.449				4046	0.497	0.500

In Figure 2, we see how the different cancer types vary in their 5-year mortality rate and share of individuals with metastasis. The least deadly cancer is breast cancer, and the deadliest cancer is lung cancer. Lung cancer is also the cancer type with the highest share of metastasis at diagnosis. Skin and breast cancer have the lowest share of metastasis. This is probably due to the fact that skin cancer is relatively easy to detect, and women are extensively screened for breast cancer, hence diagnosing many at an early stage before metastasis.

In Figure 3 we see how mortality and metastasis varies over the months of diagnosis. July is a month during which many Norwegians have vacation leave and where hospital activity is lower than other months. We see that individuals who are diagnosed in July have higher mortality rates than individuals who are diagnosed in other months. A chi-squared test of equal rates of metastasis in each month has a p-value of less than 0.001, and it is therefore unlikely that the sample came from a population in which the 5-year mortality rate is statistically independent from the month of diagnosis. Patients diagnosed in August and July seem to have higher mortality compared to patients diagnosed in other months of the year.

Figure 2: **Cancer types.**

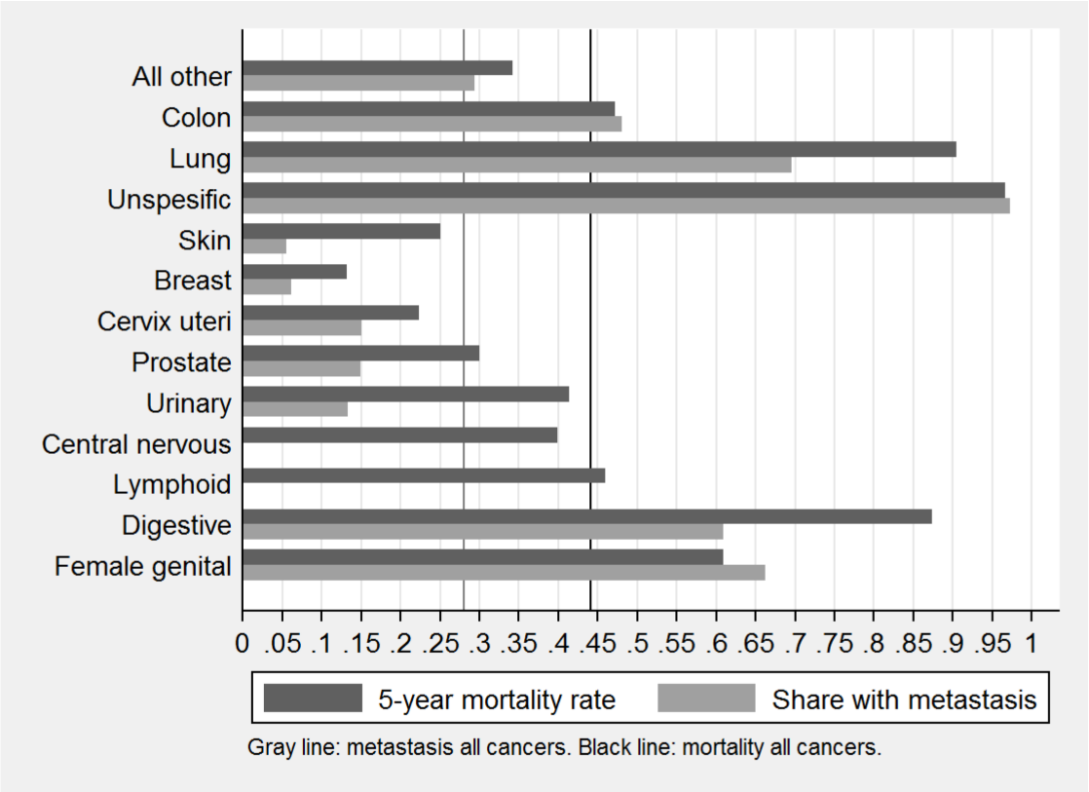
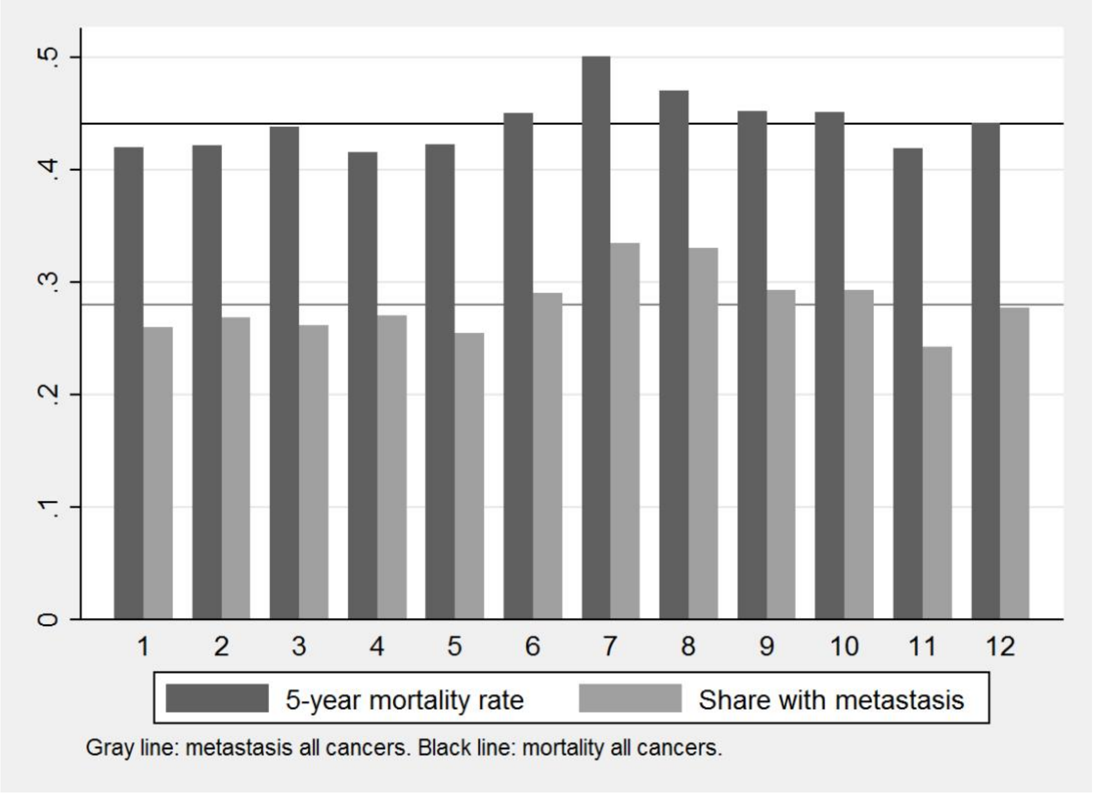


Figure 3: **Metastasis and 5-year mortality by month.**

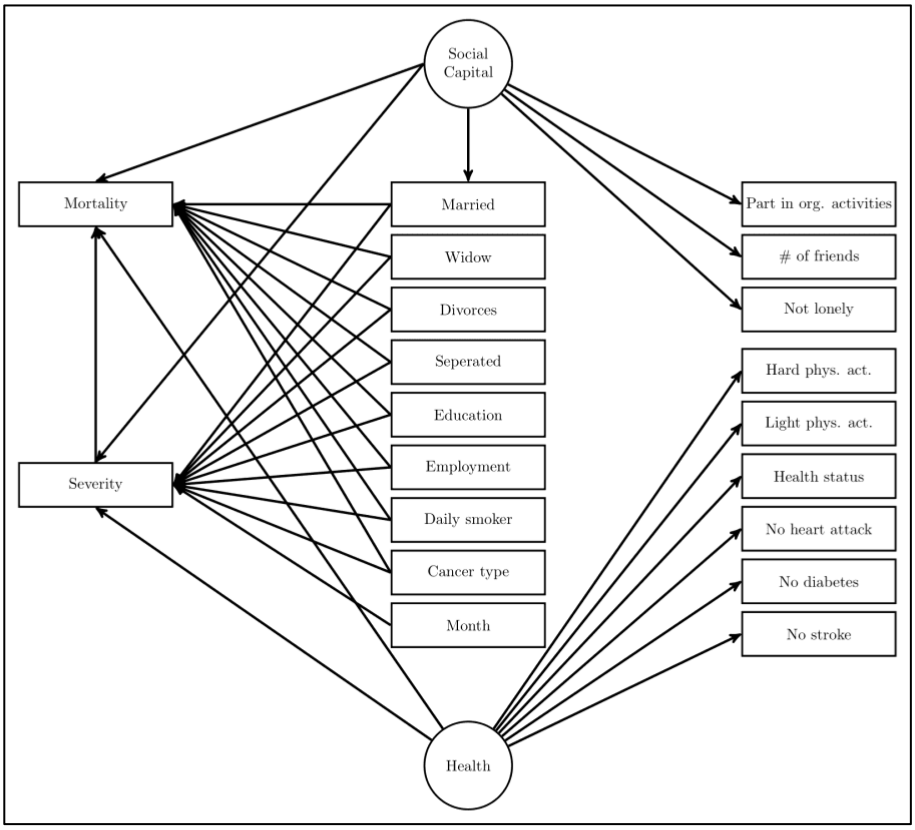


3. Model specification

To estimate a conditional statistical association between social capital on cancer severity and mortality, we specify a generalized structural equation (GSE) model. The term *generalized* refers to the fact that the link function between dependent and independent variables are not necessarily a linear function, as it could, for instance, be a sigmoid shaped function. The GSE framework also opens the possibility to include latent variables, both as dependent and independent variables. A latent variable is most generally defined as a variable where there is no sample realization (Bollen 2002). *Social capital* and *Health* are two such examples, as they are abstract constructs, which can only be measured indirectly and not perfectly by other variables in our sample. In other words, we can only infer their existence and consequences of these constructs through observed indicator variables. We specify the statistical relationship between latent variables and observed variables in a measurement model, essentially creating an index of each of our latent variables while letting the weights of each observable variable constructing the index be determined by the data. As described by Skrondal and Rabe-Hesketh (2004), GSE modeling is suitable when the variables of interest are not measured perfectly.

Our GSE model consists of two parts: a measurement model, where we identify and measure the latent variables *Social Capital* and *Health*; and a structural part where we specify two regression equations. In our regression equations, our dependent variables are cancer severity, as indicated by the presence of metastasis, and 5-year mortality. Figure 4 gives a graphic representation of our GSE.

Figure 4: Conceptual graph of our GSE model.



Health and *Social Capital* are latent variables (ovals) measured by other observed variables (rectangles). The arrows indicate how the variables are related in our model. *Health* and *Social Capital* are also non-directional related (correlated).

Figure 5: Measurement model is the part of the GSE model in which we measure the latent variables.

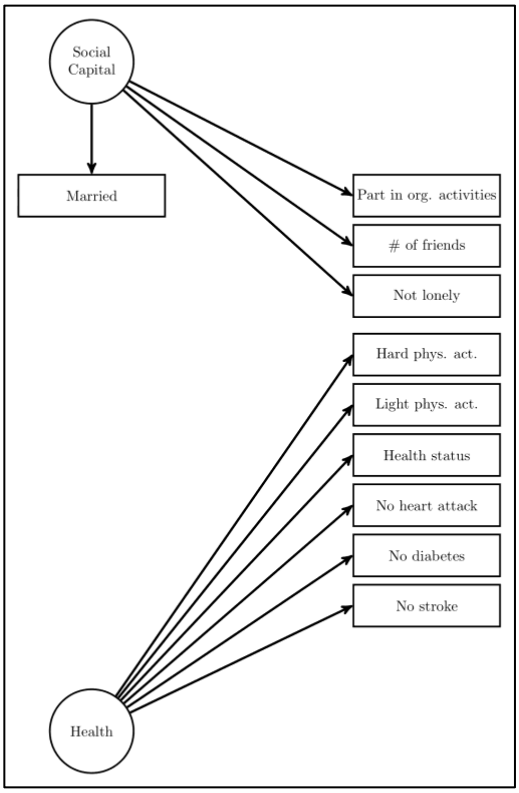
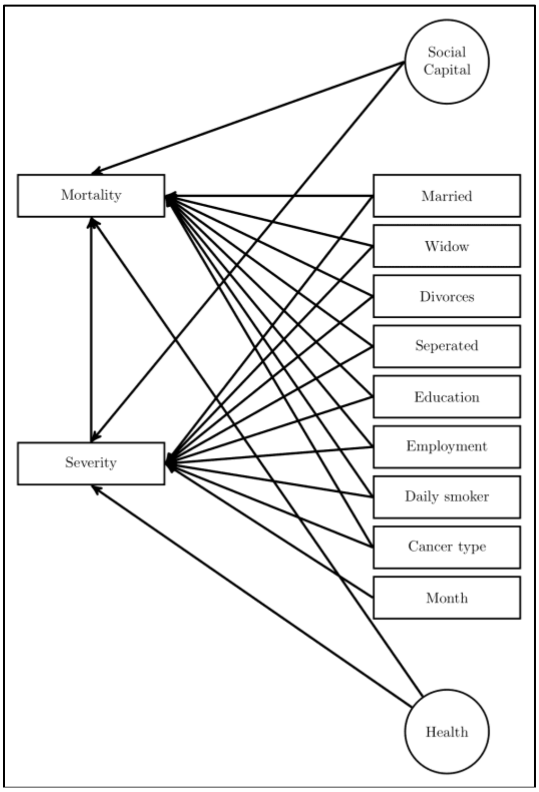


Figure 6: Structural model describes our regression equations.



3.1. Measurement model

The measurement model is technically equivalent to a two-parameter item response theory (IRT) model known from psychometrics, also referred to as a generalized confirmatory factor analysis model. Variables coded with more than two values are modelled as ordered categorical variables, since the questionnaire answers are of the form as follows: "*poor*"; "*fair*"; "*good*"; and "*excellent*". The relationship between the latent constructs (*Latent)* and the observed item-variables (x_i) in an ordered model can be described by the equation,

$$\Pr(x_i = j|Latent) = \frac{1}{1+\exp\,(-\alpha_{i,j}+\beta_i(Latent))} - \frac{1}{1+\exp\left(-\alpha_{i,j-1}+\beta_i(Latent)\right)} \tag{1}$$

In (1) we are specifying and ordered logit model, since we use the logistic function. For binary variables the relationship between the latent constructs and the observed variables are given by,

$$\Pr(x_i = 1|Latent) = \frac{\exp\,(\alpha_i + \beta_i(Latent))}{1 + \exp\,(\alpha_i + \beta_i(Latent))} \tag{2}$$

In (1) and (2) the latent variables can either be *Social Capital* or *Health*. The latent variables do not have a given scale or metric so the variances of the latent variables are constrained to 1.

$$Latent \sim N(0,1)$$

i.e., the latent variables are standard normal variances with some correlation. The endogenous variables are the observed indicator variables, and the parameters to be estimated are the α_i's and β_i's.

Equation (1) and (2) only give a general specification of the latent variables. Although each latent construct is modelled in the same general way, they are measured by two separate sets of observed variables, as shown in Figure 5 and 6. We chose the two sets of variables such that the correlation between the latent constructs is low.

3.2. Measurement items

We use four observed variables as items to measure *Social Capital*. Less institutionalized relationships we interpret as friends, and the aggregate actual resource is the number of friends. A more institutionalized relationship would be marriage and, for instance, people one knows as a result of participating in organized activities. It is natural to believe that by having high social capital, one is less likely to experience loneliness, since people with a high level of social capital have more friends. Importantly, social capital is something different from socioeconomic status and human capital, therefore we do not use employment status and/or education as items for measuring social capital. Health is perhaps more easily understood as a latent variable, as it is a common abstract concept, and most people understand that there is not a single measurement of health. In this paper, we use health condition prior to cancer diagnosis (stroke, diabetes, heart attack, and self-reported health status) as items. From our data, the social capital and health variable are defined by four items each, given in Table 3.

Table 3: Measurement Items.

Item / Indicator variable	Latent Variable	Explanation	Data Type
PARTICIPATES IN ORGANIZED ACTIVITIES	*Social Capital*	High *Social Capital* increases the likelihood to participate in org. activities	Ordered Categorical
NUMBER OF FRIENDS	*Social Capital*	High *Social Capital* increases the likelihood that the person has many close friends	Ordered Categorical
NOT LONELY	*Social Capital*	High *Social Capital* increases the risk of being not lonely	Ordered Categorical
MARRIED	*Social Capital*	High Social Capital increases the risk of being married	Binary
NO STROKE	*Health*	Better *Health* increases the likelihood of no stroke	Binary
NO DIABETES	*Health*	Better *Health* increases the likelihood for no diabetes	Binary
NO HEART ATTACK	*Health*	Better *Health* increases the likelihood of no heart attack	Binary
SELF-REPORTED HEALTH STATUS	*Health*	Better *Health* increases the likelihood for reporting good health	Ordered Categorical
LIGHT PHYSICAL ACTIVITY	*Health*	Better *Health* increases the likelihood for reporting light (not sweating or out of breath)	Ordered Categorical
HARD PHYSICAL ACTIVITY	*Health*	Better *Health* increases the likelihood for reporting hard (sweating/out of breath)	Ordered Categorical

3.3. Structural model

The relations in the structural model is specified in Equations 3 and 4. These are non-linear relations. However, the parameter estimates of 3 and 4 can be interpreted in the same way as parameter estimates from an ordinary probit regression.

$$P\big(Y_{1j} = 1 \big| x_{ij}, SocCap_j, Health_j\big) = \phi(\beta_{1i}x_{ij} + \eta_{1_s}SocCap_j + \eta_{1_h}Health_j) \tag{3}$$

$$P\big(Y_{2j} = 1 \big| y_{1j}, x_{ij}, SocCap_j, Health_j\big) = \phi(\beta_{2i}x_{ij} + \gamma y_{1j} + \eta_{2_s}SocCap_j + \eta_{2_h}Health_j) \tag{4}$$

In our specific model, Y_{1j} is cancer metastasis for person *j*, and Y_{2j} is 5-year mortality for person *j*. Our key research questions translate to whether the coefficients η_{1s}, η_{1h}, η_{2s} and η_{2h}, have low p-values, given the null hypothesis that the coefficient is zero. The raw estimates of the coefficients η_{1s}, η_{1h}, η_{2s} and η_{2h} are difficult to interpret. We will calculate the marginal effects of statistically significant variables to ease interpretation. We also note that the variable married is both a regressor in the structural equations and a measurement variable for the latent variable *Social Capital*. Married is therefore a *mediator* between *Social Capital* and severity and *Social Capital* and mortality. Keeping the variable married in the regression equations reflects that our aim is to estimate of the *direct effect* of social capital on severity and mortality. The rationale is that being married might have direct effect that does not operate via Social Capital, since having a spouse can be helpful in uncovering some types of cancer (e.g. skin cancer) or assist with life style changes. Since our interest is in how social capital can favor individuals in the healthcare system, we control for this “spouse effect'' in the structural model.

3.4. Identification of the regression equations and causality

The existence of unobserved confounding variables can never be completely ruled out when analyzing data. However, we attempted to assess the bias in our model by estimating a correlation between the regression equations. To do this, we estimate a standard bivariate probit regression model without the latent variables, but where we directly include the measurement variables in the regression equations. The correlation between the two equations is low and not statistically significant. A low correlation indicates that the regression equations are independent and, therefore, it is not likely that there are any critical biasing unmeasured confounding variables between cancer severity and mortality.

Since month of year is associated with cancer severity, and we may argue that month of year is not directly associated with mortality, month of year is a potential instrumental variable (IV) in the model. However, it is unclear how strong an IV month of year is, and therefore how credible it is, so we do not pursue this idea any further.

A rather esoteric method of model identification is to obtain identification by means of functional form. Given the assumption that mortality and cancer metastasis are truly bivariate variables, the log likelihood function is a function of independent probabilities, which are restricted to sum to one, and this restriction provides parameter identification in our case. We refer to Wilde (2000) and Greene (2012, 778-789) for more details. Identification by means of functional form is regarded as unconventional and has not been commonly applied in recent years.

Model identification does not necessarily imply that we are estimating causal effects. For our estimates to represent causal effects, we must assume that our model is describing something close to the true data generating process, which is a strong assumption and unlikely to hold in reality. Nevertheless, it is important to note that we are unlikely to have a problem of selection bias in our sample, since individuals were selected into the CONOR study before any knowledge of a cancer diagnosis. Still, the effect estimates we present in the next section should be interpreted cautiously with regards to causality, and the default interpretation is conditional statistical associations.

4. Results

In Table 4 we have reported the relations between items (Dependent variables) and latent constructs. We see that absence of loneliness, having more friends, and being married are factors that loads positively on *Social Capital*. As expected, better self-assessed health loads positively on the latent variable *Health*. Furthermore, light physical activity, and absence of stroke, diabetes, and heart attack also have positive factor loadings on *Health*. The results indicate that the two latent variables *Social Capital* and *Health* are positively correlated, with a Pearson correlation estimate of 0.29. We therefore include both the latent variables *Health* and *Social Capital* in both of the structural regression equations.

Table 4: Results from maximum likelihood estimation of measurement model.

Latent variable: *Social Capital*				
Dependent variable			**Estimate**	**Std. Err.**
NOT LONELY		Slope	3.78	(0.777)***
		Const1	-9.37	(1.689)
		Const2	-6.73	(1.001)
		Const3	-3.53	(0.723)
ACTIVITIES		Slope	0.15	(0.046)***
		Const	-1.91	(0.038)***
FRIENDS		Slope	0.41	(0.038)***
	5 to 9	Const1	-0.78	(0.028)
	10 to 14	Const2	0.50	(0.027)
	15 to 19	Const3	1.87	(0.038)
	20 to 24	Const4	2.36	(0.045)
	> 24	Const5	3.41	(0.070)
MARRIED		Slope	1.08	(0.074)***
		Const	0.82	(0.033)***
Latent variable: *Health*				
Dependent variable			**Estimate**	**Std. Err.**
HEALTH STATUS		Slope	0.98	(0.043)***
	poor	Const1	-3.92	(0.073)
	good	Const2	-0.61	(0.027)
	very good	Const3	2.64	(0.048)
NO STROKE		Slope	0.79	(0.075)***
		Const	3.36	(0.074)***
NO DIABETES		Slope	0.73	(0.068)***
		Const	3.05	(0.063)***
NO HEART ATTACK		Slope	0.64	(0.059)***
		Const	2.67	(0.051)***
LIGHT PHYS. ACTIVITIES		slope	1.33	(0.056)***
		Const1	-2.73	(0.059)
		Const2	-1.56	(0.041)
		Const3	0.28	(0.030)
HARD PHYS. ACTIVITIES		Slope	1.92	(0.105)
		Const1	-0.03	(0.037)
		Const2	1.47	(0.061)
		Const3	3.30	(0.109)
corr(*Social Capital, Health*): 0.29 (0.023)***				

*Significant at the 10 % level. **Significant at the 5 % level, ***Significant at the 1 % level

The estimated coefficients for our structural model are presented in Table 5. For the latent variables Social Capital and Health, the estimated coefficients are the z-scores on the probability that a cancer is metastatic or that the patient dies within 5 years after diagnosis as Social Capital or Health changes by one standard deviation. We see that the latent variable Health has negative and statistically significant effect on cancer mortality, while we do not find a significant association with cancer severity metastasis.

Table 5: Results from maximum likelihood estimation of the structural model.

	Dep. Var. Metastasis		**Dep. Var. 5-year mortality**	
Regressors	**Estimate**	**Standard Error**	**Estimate**	**Standard Error**
Social Capital	0.01	(0.035)	-0.05	(0.035)
Health	0.01	(0.026)	-0.14	(0.026)***
CANCER METASTASIS	-	-	1.17	(0.044)***
A SECOND CANCER DIAGNOSIS	-	-	0.26	(0.057)***
AGE	-0.01	(0.002)***	0.03	(0.002)***
MALE	0.03	(0.044)	0.18	(0.043)***
MARTIAL STATUS				
UNMARRIED	0.01	(0.066)	0.18	(0.064)***
WIDOW	0.03	(0.063)	0.03	(0.061)
DIVORCED	-0.12	(0.066)*	-0.03	(0.065)
SEPARATED	-0.07	(0.166)	-0.09	(0.160)
HIGHER EDUCATION	-0.004	(0.047)	-0.09	(0.045)**
DAILY SMOKER	0.03	(0.039)	0.12	(0.038)***
EMPLOYMENT	-0.06	(0.049)	-0.12	(0.047)**
MONTH OF DIAGNOSIS				
JANUARY	0.10	(0.082)	-	-
FEBRUARY	0.12	(0.084)	-	-
MARCH	0.05	(0.083)	-	-
APRIL	0.13	(0.083)	-	-
MAY	0.04	(0.082)	-	-
JUNE	0.20	(0.080)**	-	-
JULY	0.16	(0.084)*	-	-
AUGUST	0.24	(0.084)***	-	-
SEPTEMBER	0.15	(0.081)*	-	-
OCTOBER	0.21	(0.080)***	-	-
DECEMBER	0.11	(0.083)	-	-
Dummy for type of cancer	yes		yes	
CONSTANT	0.77	(0.162)***	-1.48	(0.160)***
N	8496		8496	

*Significant at the 10 % level. **Significant at the 5 % level, ***Significant at the 1 % level

Based on our estimation results, we cannot conclude if cancer mortality or metastasis is associated with *Social Capital*. An important note is that almost all predictive power for cancer metastasis lies within the type of cancer a person gets. For 5-year mortality, metastasis at the time of diagnosis is in addition to the type of cancer explaining most of the differences. Surprisingly, we see that an older age is associated with a lower likelihood of cancer metastasis, however, the estimate is low and might not be of clinical importance. It might be because that as people get older, they are more in contact with the healthcare system. However, being older significantly increases the 5-year mortality. There is also a statistically significant association from the months of June, August, and October on the likelihood of cancer metastasis, which we expected from the graph above. One can speculate that this association has a connection with major holidays in Norway. However, a more thorough study and a new independent sample is required to address this question. For 5-year mortality, being unmarried (relative to being married or having a partner) and being a daily smoker increases mortality significantly, while higher education and employment significantly reduces mortality. This is as expected, as we know that higher education has long been known to be associated with a longer life expectancy.

4.1 Average treatment effects

The coefficients for a binary variable in the bivariate probit model are difficult to interpret directly. In our case we compute the average treatment effect from binary variable i as the average difference in predicted probability, such predictions are also called marginal effects.

$$\frac{1}{N}\sum_{n=1}^{N}\left(P_n\left(x'_{-i}\widehat{\beta_{-\iota}}+\widehat{\beta}_\iota\right)-P_n\left(x'_{-i}\widehat{\beta_{-\iota}}\right)\right).$$

P_n is the predicted probability when the other variables are held at their observed values. For the ATE, we select the variables from Table 5 that have a significant level 5 % or lower. From Table 6 we see that it is more likely to be diagnosed with a metastasis in patients diagnosed in June, August, or October, with percentages of 4.97%, 6.01% and 5.02%, respectively. For 5-year mortality, being diagnosed with a metastasis, increases the mortality by 34.72%.

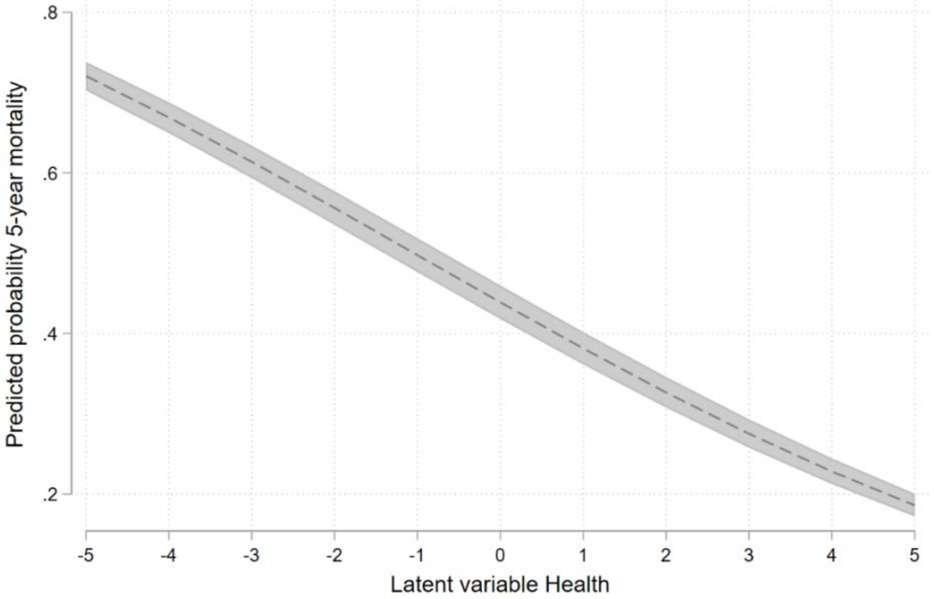
Table 6: Average Treatment Effects.

Regressors	**Metastasis**	**5-year mortality**
MONTH OF DIAGNOSIS		
[DIAGNOSED IN JUNE]	4.97 %	-
[DIAGNOSED IN AUGUST]	6.01 %	-
[DIAGNOSED IN OCTOBER]	5.02 %	-
CANCER METASTASIS	-	34.72 %
A SECOND CANCER DIAGNOSIS	-	6.91 %
MALE	-	4.71%
EMPLOYMENT	-	- 3.04 %
UNMARRIED	-	4.35 %
DAILY SMOKER	-	3.16 %
HIGHER EDUCATION		- 2.30 %

The graph in Figure 7 shows the conditional effect of the latent variable *Health* on 5-year mortality. The estimated effect depends on keeping *Social Capital* at zero and the mean of the equation's predicted “linear'' portion describing 5-year mortality. We also provide a confidence interval based on the model output. However, calculating the “correct'' confidence intervals for effects derived from latent variables in a non-linear GSEM presents a non-trivial challenge analytically and computationally. Assessing the balance between cost and gains (in the form of quantified uncertainty), we consider that the cost outweighs the gains for such computations, and we do not pursue it any further. Therefore, the reader must remember that the confidence interval in Figure 7 understates the true model uncertainty.

As an example, a person with *Health* level two standard deviations below average, has, on average, a 5-year mortality probability around 0.5. On the opposite side of the scale, we find the person with *Health* two standard deviations above the average, has on average a 5-year mortality probability roughly around 0.3.

Figure 7: Estimated effect of *Health* on mortality.



5. Discussion

To obtain information about the individuals prior to diagnosis, we used information from several health surveys. For some individuals, the background information was collected up to six or seven years before the time of cancer diagnosis. There have been changes in some of these factors, such as self-reported health and comorbidity status, as well as social capital. Still, we believe it is reasonable to assume that information about an individual at one point in time provides relevance for the future individual characteristics. Furthermore, some variables, such as education, are likely not to change over the years.

We are able to account for several individual characteristics simultaneously, while often only a subgroup of variables has been included in other studies. We simultaneously account for metastasis at the time of diagnosis and mortality to identify the marginal effect of covariates on mortality. However, measurement for *Health* are all self-reported measures, which might provide bias. Further, *Social Capital* includes only individual level factors, hence we are not able to adjust for characteristics at the community level.

We have included all-cause mortality, not separating between cancer-specific mortality, and death due to other causes. Within a five-year perspective, the majority of deaths were cancer related, hence related to either metastasis at the time of diagnosis, or recurrence within the five-year perspective. We employ a binary measure to indicate the presence or absence of metastasis, and this measure serves as a rough measure of cancer severity. There are several other methods to identify severity, for instance by applying the TNM staging system. Including additional dimensions of severity could potentially have provided more detailed knowledge on the relationship between social capital and health on cancer severity and mortality. For instance, whether a metastasis is in the same organ or spread to another organ, which would be identified with the TNM staging system, influences the prognosis.

Breast and cervical cancer have been subject to organized screening programs for several decades. In our study, we do not have information about whether a cancer was detected through screening or due to symptoms. Nevertheless, we adjust for the type of cancer (including breast and cervical cancer) in our analysis. As a result, our effect estimates may reflect the impact of screening programs: Both breast and cervical cancer have a significant lower likelihood of being diagnosed with metastasis.

6. Conclusion

We revisit the question of how cancer mortality relates to an individual's general health status and social capital. Our contribution is novel in two ways: first, we use longitudinal data to collect individuals' background information before cancer diagnosis, and second, we identify the effects of an individual's general health and social capital by defining these concepts as latent variables that are indirectly observable. Our empirical approach allows us to quantify the separate effect of *Health* and *Social Capital* on cancer mortality and cancer severity at the time of diagnosis. We find that health has a negative and statistically significant effect on cancer mortality, while we do not find a significant effect on the probability of cancer metastasis at the time of diagnosis. We cannot reject the null hypothesis that cancer mortality and the probability of cancer metastasis is unaffected by *Social Capital*. An interesting finding is that being diagnosed with cancer between June and October is associated with a significant increase in the probability of having more severe cancer at the time of diagnosis.

Our results confirm findings from Kravdal (2000 and 2001), where he identifies a significant effect of marital status and education on mortality. In this study, we take advantage of a rich dataset, including several individual characteristics that are expected to have an impact on access to healthcare and survival. From the provider's perspective, the analysis reveals that metastasis at the time of diagnosis is distributed differently between months. Further research should pay closer attention to this finding, as we know that metastasis increases mortality. The introduction of cancer care plans (*Pakkeforløp for kreft*), which provide specific deadlines for diagnostics, start of treatment, and follow-up, has the potential to reduce this gap.

Acknowledgements

We thank anonymous reviewers for valuable comments and suggestions. The project was funded by NORCHER (Norwegian Centre for Health Services Research), funded by the Norwegian Research Council grant no 296114.

7. References

Aamodt, Geir, Anne Johanne Søgaard, Øyvind Næss, Anne Cathrine Beckstrøm, og Sven Ove Samuelsen. 2015. *Cohort of Norway (CONOR) - Forskningspotensial, design og representativitet.* Folkehelseinstituttet (FHI).

Auvinen, A., S. Karjalainen, and E. Pukkala. 1995. "Social Class and Cancer Patient Survival in Finland." *American Journal of Epidemiology*, 1089-1102.

Auvinen, Anssi. 1992. "Social Class and Colon Cancer Survival." *Cancer*, 402-409.

Bollen, Kenneth A. 2002. «Latent Variables in Psychology and the Social Sciences.» *Annual Review of Psychology*, 603634. doi:10.1146/annurev.psych.53.100901.135239.

Fredriksen, B. L., M. Osler, H. Harling, and T. Jørgensen. 2008. "Social inequalities in stage at diagnosis of rectal but not in colonic cancer: a nationwide study." *British Journal of Cancer*, 668-673.

Glaeser, Edward L., David Laibson, and Bruce Sacerdote. 2002. "An Economic Approach to Social Capital." *The Economic Journal*, 437-458.

Goodwin, J. S., W. C. Hunt, C. R. Key, and J. M. Samet. 1987. "The effect of martial status on stage, treatment, and survival cancer patients." *Journal of American Medical Association*, 3125 - 3130.

Greene, William H. 2012. *Econometric Analysis.* 7. Pearson.

Grossman, Michael. 1972. "On the Concept of Health Capital and the Demand for Health." *Journal of Political Economy*, 223-255.

Holt-Lunstad, Julianne, Timothy B. Smith, and J. Bradley Layton. 2010. "Social Relationships and Mortality Risk: A Meta-analytic Review." *PLoS Med*, 7 ed.

Islam, M. Kamrul, Juan Merlo, Ichiro Kawachi, Martin Lindström, and Ulf-G Gerdtham. 2006. "Social capital and health: Does egalitariansim matter? A literature review." *International Journal og Equity in Health.*

Johansen, C., G. Schou, H. Soll-Johanning, A. Mellemgaard, and E. Lynge. 1996. "Influence of martial status on survival from colon and rectal cancer in Denmark." *British Journal of Cancer*, 985-988.

Kravdal, Ø. 2000. "Social inequalities in cancer survival." *Population Studies*, 1-28.

—. 2001. "The impact of martial status on cancer survival." *Social Science and Medicine*, 357-368.

Paxton, Pamela. 1999. «Is Social Capital Declining in the United.» *American Journal of Sociology*, 105. utg.: 88-127.

Scheffler, R.M., T.T. Brown, L and Kawachi, I Syme, I Tolstykh, and C Iribarren. 2008. "Community-level social capital and recurrence of acute coronary syndrome." *Social Science and Medicine*, 1603-1613.

Skrondal, Anders, and Sophia Rabe-Hesketh. 2004. *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models.* Chapman & Hall/CRC.

StataCorp LLC. 2017. *STATA Item Response Theory Reference Manual Release 15.* College Station, Texas: Stata Press Publication.

The norwegian ministry of health and care services. 2007. *National strategy to reduce social inequalities in health.* Report No. 20 (2006–2007) to the Storting, Oslo: The norwegian ministry of health and care services.

Villingshøj, M., L. Ross, B. Thomsen, and C Johansen. 2006. "Does martial status and altered contact with the social network predict colorectal cancer survival?" *European Journal of Cancer*, 3022-3027.

Wilde, Joachim. 2000. "Identification of Multiple Equation Probit Models with Endogenous Regressors." *Economics Letters*, April: 309–312.

Wooldridge, Jeffrey M. 2010. *Econometric Analysis of Cross Section and Panel Data.* 2. The MIT Press.

Appendix

A.1 Descriptive Statistics

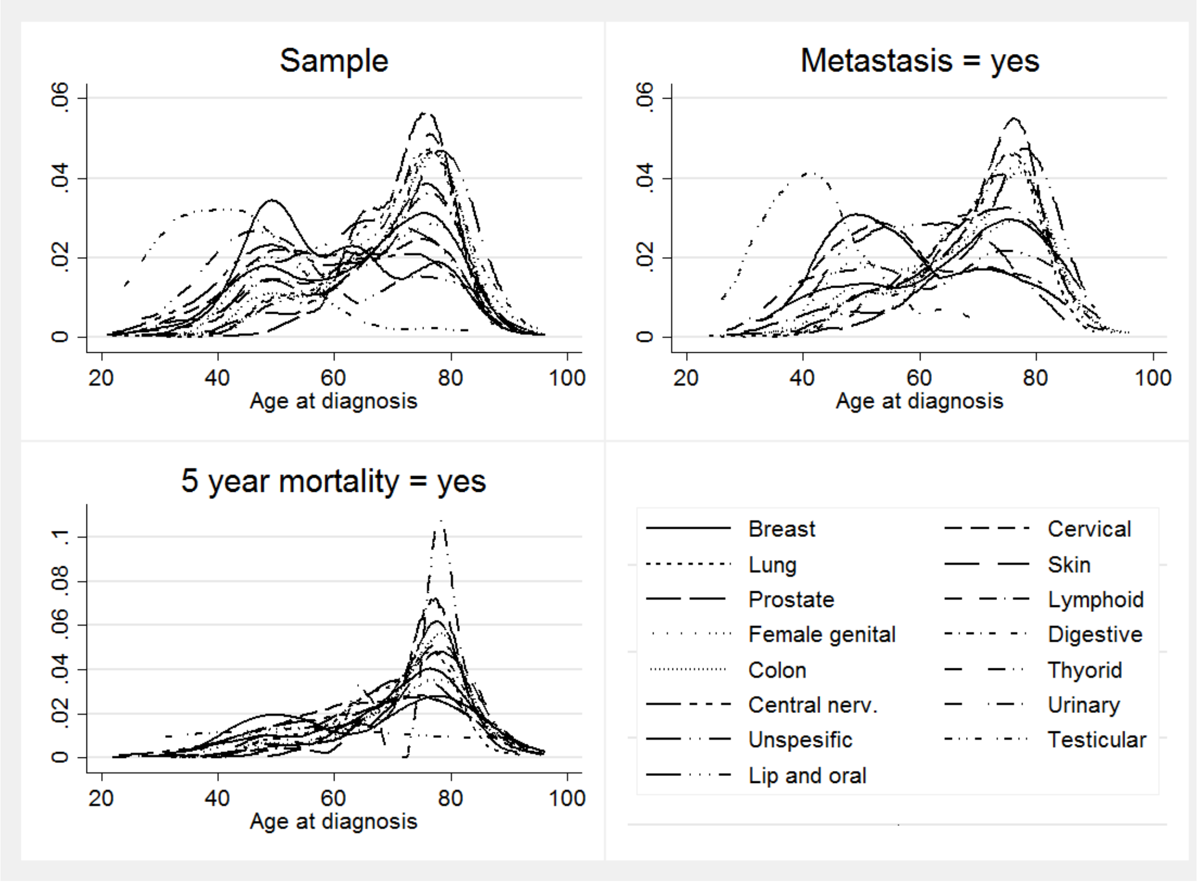
Table A1: Comparison of counties in CONOR compared to the general Norwegian population according to selected measures.

Characteristics	**Norway**	**Troms and Finnmark**	**Trøndelag (Nord-Trøndelag)**	**Vestland (Hordaland)**	**Oslo**
Life expectancy*					
Female	81.3	80.7	81.4	82.4	80.2
Male	75.7	74.2	76.1	76.5	74.6
Proportion men (1999)	0.49	0.51	0.50	0.50	0.48
Proportion above 80 (1999)	0.042	0.035	0.043	0.045	0.047
Cancer rate (unadj.)					
Female	549.7	528.7	531.1	556.4	555.3
Male	706.0	706.0	715.3	762.4	688.9
Proportion with higher education (2021)	0.36	0.26	0.34	0.30	0.51
Median household income (2016) NOK	498000	493000	499000	517000	448000
Proportion living alone (45 years + in 2017)	0.25	0.26	0.25	0.24	33.7
Proportion smoking (female – 2013-2017)	0.059	0.082	0.045	0.044	0.028

Table A2: Descriptive statistics of the study sample.

Variable	Obs.	Mean	Std. Dev.	Min	Max
METASTASIS	9179	0.2798	0.45	0	1
5-YEAR MORTALITY	9179			0	1
SECOND CANCER	9179	0.0788	0.27	0	1
SMOKING	9059	0.2980	0.46	0	1
HEALTH STATUS	9067	2.6880	0.68	1	4
EMPLOYMENT	9179	0.3244	0.47	0	1
ACTIVITIES	9179	0.1301	0.34	0	1
HAS ENOUGH FRIENDS	9179	0.5729	0.49	0	1
NUMBER OF FRIENDS	6427	1.3325	1.31	0	5
NOT LONELY	6657	2.7191	0.60	0	3
MARRIED	9117	0.6609	0.47	0	1
NO STROLE	8974	0.9564	0.20	0	1
NO DIABETES	9007	0.9437	0.23	0	1
NO HEART ATTACK	9002	0.9242	0.26	0	1
LIGHT PHYSICAL ACTIVITY	8082	3.1142	0.99	1	4
HARD PHYSICAL ACTIVITY	7177	1.9149	1.04	1	4
MALE	9179	0.5647	0.50	0	1
AGE	9179	67.5270	12.81	21	96
DIAGNOSIS MONTH					
JANUARY	9179	0.0824	0.27	0	1
FEBRUARY	9179	0.0789	0.27	0	1
MARCH	9179	0.0854	0.28	0	1
APRIL	9179	0.0775	0.27	0	1
MAY	9179	0.0862	0.28	0	1
JUNE	9179	0.0921	0.29	0	1
JULY	9179	0.0707	0.26	0	1
AUGUST	9179	0.0748	0.26	0	1
SEPTEMBER	9179	0.0875	0.28	0	1
OCTOBER	9179	0.0921	0.29	0	1
NOVEMBER	9179	0.0936	0.29	0	1
DECEMBER	9179	0.0790	0.27	0	1
MARTIAL STATUS					
MARRIED	9117	0.6609	0.47	0	1
UNMARRIED	9117	0.1007	0.30	0	1
WIDOW	9117	0.1398	0.35	0	1
DIVORCED	9117	0.0871	0.28	0	1
SEPARATED	9117	0.0115	0.11	0	1
CANCER DIAGNOSIS					
COLON	9179	0.1477	0.35	0	1
LUNG	9179	0.0904	0.29	0	1
UNSPESIFIC	9179	0.0159	0.13	0	1
SKIN	9179	0.0777	0.27	0	1
BREAST	9179	0.1190	0.32	0	1
CERVIX UTEROUS	9179	0.0312	0.17	0	1
PROSTATE	9179	0.1767	0.38	0	1
URINARY	9179	0.0751	0.26	0	1
CENTRAL NERVOUS SYSTEM	9179	0.0341	0.18	0	1
LYMPHOID	9179	0.0770	0.27	0	1
DIGESTIVE	9179	0.0733	0.26	0	1
FEMALE GENITALS	9179	0.0248	0.16	0	1
THYORID	9179	0.0077	0.09	0	1
TESTICULAR	9179	0.0074	0.09	0	1
LIP AND ORAL	9179	0.0146	0.12	0	1

Figure A1: Distribution of age at time of diagnosis for different types of cancer.



A.2 Measurement Model

The measurement model is equivalent to and IRT-model. IRT-models have historically been used in education and psychology to measure ability and other cognitive constructs. Health has more recently been the target of IRT-modelling. More generally, IRT-models can be referred to as generalized factor analysis. Standard factor analysis models a linear relationship between the latent variable and its items, by generalized we mean that this relationship is not necessarily linear. To assess the measurement model, we use graphs and curves.

Characteristic curves describe an item score's probability as a latent variable's function. The steeper the characteristic curve, the better the item is to distinguish between low and high values of the latent trait. The characteristic curves of *Social Capital* are given in Figure A2. In Figure A2 the curves representing the item not lonely are the steepest; hence the feelings of loneliness are the strongest item for distinguishing the latent trait level. The item *not lonely* measures *Social Capital* best when *Social Capital* has low values but is also the best measure of *Social Capital* overall.

Figure A2: BCC *Social Capital.*

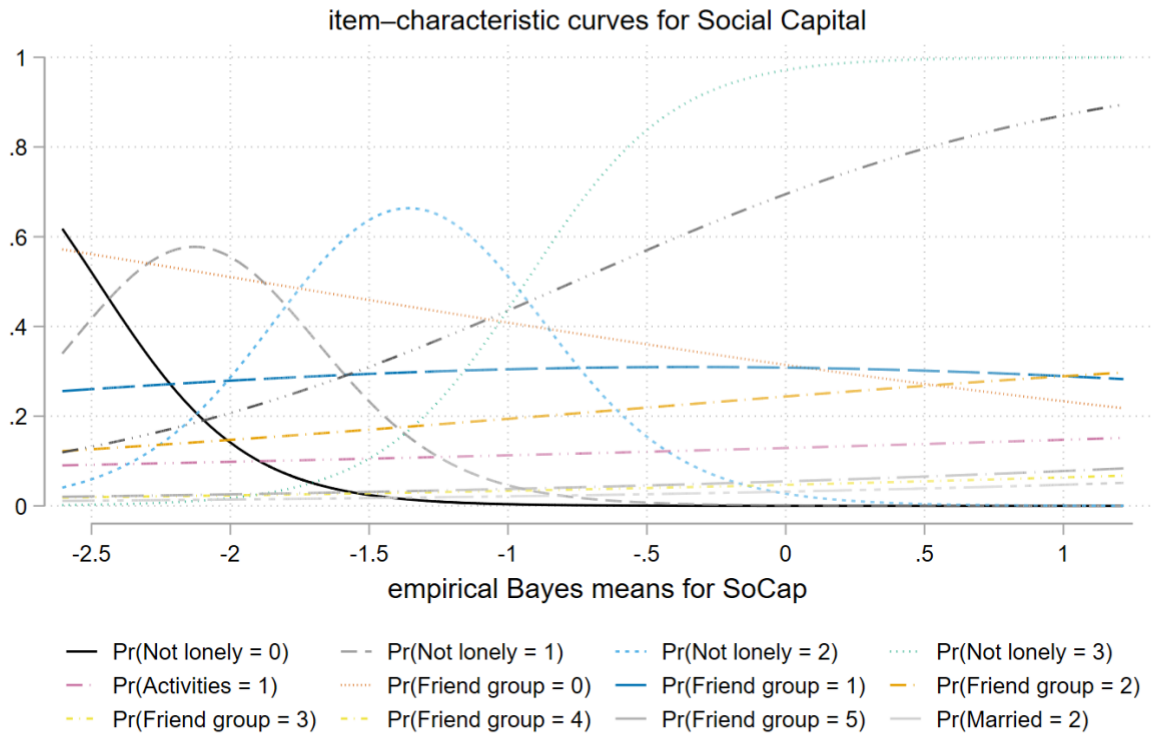
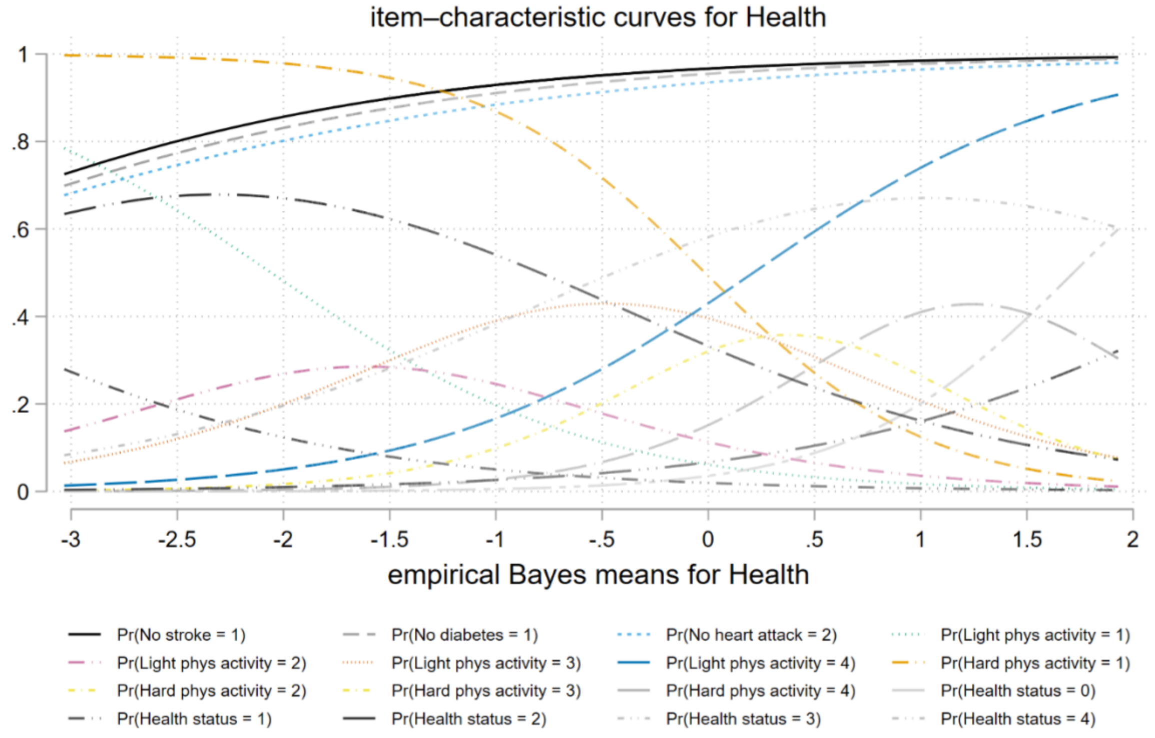


Figure A3: BCC *Health.*



The best item for measuring health is hard physical activity. The latent trait Health is evenly measured for both low and high levels of health. We can also interpret from the curves for no stroke, no diabetes, and no heart attack, that people who have had a stroke, have diabetes, or have had a heart attack have substantially lower general health than people without these conditions.

Two histograms of the predicted values for the latent variables are given in Figure A4 and Figure A5. The latent variable *Health* is symmetrical around 0. The predicted distribution of *Social Capital* is skewed or consisting of two peaks reflecting that it is well measured for low values, it also indicates the fact the social capital is more difficult to measure than health.

Figure A4: Empirical Bayes means: *Social Capital.*

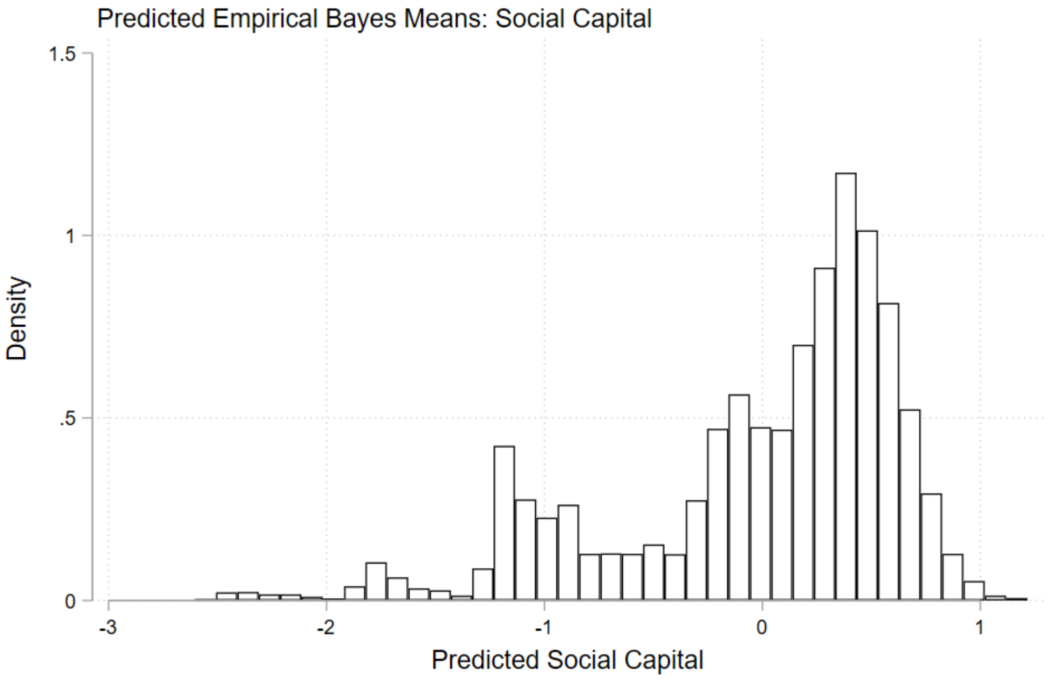
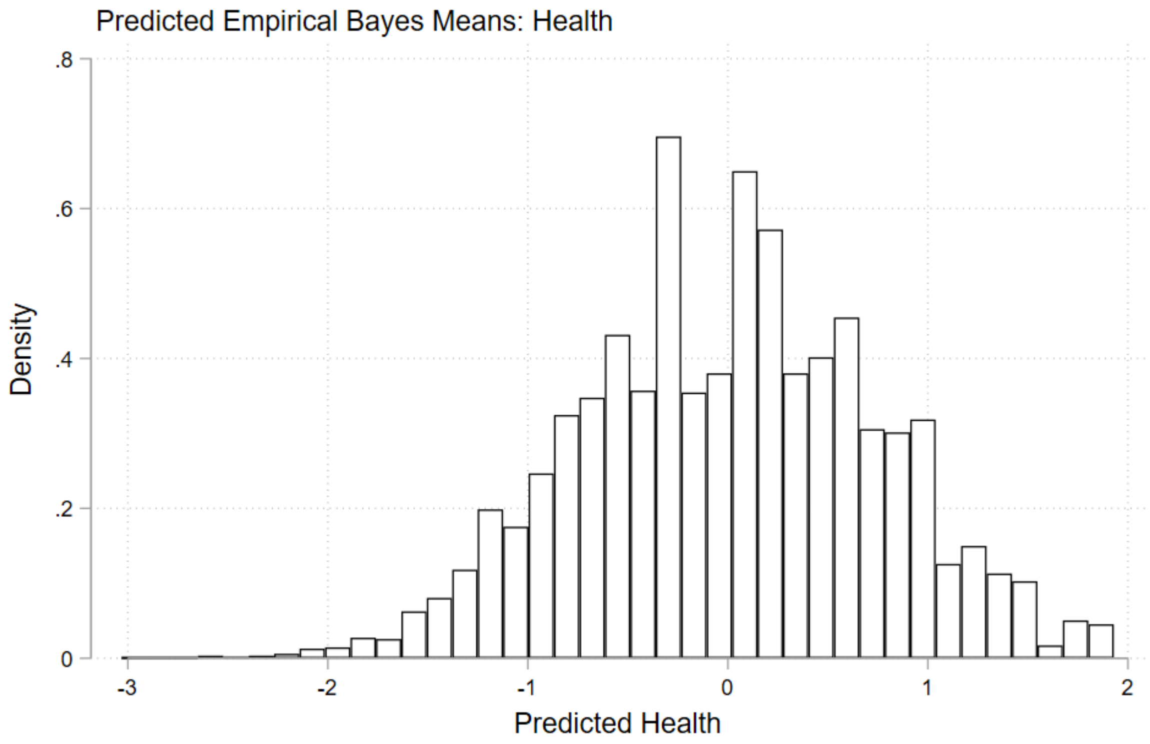


Figure A5: Empirical Bayes means: *Health*.


Predicted Empirical Bayes Means: Health
Density
Predicted Health