SHARE
Survey of Health, Ageing and Retirement in Europe
50+ in Europe

The 4th International
SHARE User Conference
HEC Business School

SZKOŁA GŁÓWNA HANDLOWA W WARSZAWIE

PREDICTION OF PROBABLITY OF HOSPITALIZATION AMONG PEOPLE AGED 50+

Jadwiga Borucka, Joanna Romaniuk

Warsaw School of Economics
Event History and Multilevel Analysis Unit

Warsaw, Poland

Liège, Belgium
November 28-29, 2013

Presentation Plan

1. Introduction
2. Logistic regression model
3. Dataset used for analysis
4. Results and interpretation
5. Conclusions

INTRODUCTION

POPULATION AGEING:

- A problem that concerns all European countries nowadays.
- Group of older people is becoming relatively larger and constitutes higher and higher share of the total population.

According to the report "*The 2012 Ageing Report: Underlying Assumptions and Projection Methodologies*" issued by the European Commission and Economic Policy Committee (2011):

- *In the European Union the number of people aged 65+ is projected to grow from 87.5 million in 2010 to 150.2 million in 2050 and 152.6 million in 2060.*
- *Demographic dependency ratio (65+) for the European Union is forecasted to double from 26% in 2010 to 52.5% in 2060.*
- *The number of people aged 80+ is projected to almost triple in the European Union (jumping to 62.4 million in 2060 from 23.7 million in 2010).*

INTRODUCTION

POPULATION AGEING:

- Europe is having to deal with an ever-growing aging population → health issues have raised new and important concerns.
- Number of people who retire is constantly increasing and number of working employees per one pensioner is decreasing → burdens to the public health care system raise rapidly.
- Of particular interest are activities focused on identification of people having relatively higher risk of being hospitalized early enough to undertake some preventive actions.

AIMS:

- *Assess whether logistic regression might be found useful in analyses of risk of hospitalization among people aged 50+.*
- *Identify risk factors that increase probability of being hospitalized among analysed respondents.*

LOGISTIC REGRESSION MODEL

Logistic regression: one of the generalized linear models

Binary logistic regression: modelling conditional probability of taking one of the two possible outcomes by the dependent (binary) variable

Let's have Y – dependent response variable:
1 = success
0 = failure

Let's assign probabilities:
p = probability of success, $P(Y = 1) = p$
1 – p = probability of failure, $P(Y = 0) = 1 - p$

LOGISTIC REGRESSION MODEL

Logistic regression uses odds as the measure of probability

Odds is defined as:

$$odds = \frac{p}{1-p}$$

and – for p falling within the interval (0, 1) – odds is within the interval (0, +∞) and its logarithm (logit):

$$logit(p) = ln\frac{p}{1-p}$$

is within (-∞, +∞)

LOGISTIC REGRESSION MODEL

Conditional probability of Y taking value of 1, given the explanatory variables vector **x** is defined as follows:

$$\pi(\boldsymbol{x}) = P(Y = 1|\boldsymbol{x})$$

Using logit transformation we can obtain the logistic regression model expressed as follows:

$$logit[\pi(\boldsymbol{x})] = ln\frac{\pi(\boldsymbol{x})}{1 - \pi(\boldsymbol{x})} = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k = h(\boldsymbol{x})$$

and – after transformation – we can get the formula for estimator of the conditional probability, which is:

$$\hat{\pi}(\boldsymbol{x}) = \frac{\exp[h(\boldsymbol{x})]}{1 + \exp[h(\boldsymbol{x})]}$$

DATASET USED FOR ANALYSIS

Source of data: SHARE Wave 4 Survey

The event of interest: hospitalization during the last 12 months

Binary response variable is defined as:
Y = 1 if respondent was hospitalized during the last 12 months
Y = 0 otherwise

Variables considered as **covariates**:
Age (in years)
Household income (in euro)
Gender (1 = male, 0 = female)
Number of years during which respondent smoked cigarettes
Alcohol use at least three times a week (1 = yes, 0 = no)
Healthy diet, defined as use of dairy products, eggs, meat, fish, fruits and legumes at least three times a week (1 = yes, 0 = no)
Physical activity, defined as intense or medium physical activity at least once a week (1 = yes, 0 = no)

DATASET USED FOR ANALYSIS

Modelling process details:

- *In order to generalize results for the entire population,* ***calibrated weights*** *for individuals are being used*

- *After exclusion of missing/erroneous observation the final dataset contains* ***19 478 units***

- *In order to verify prediction ability of the model, the final dataset was split into* ***estimation dataset*** *(15 000 observations) and* ***validation dataset*** *(4 478 observations)*

- *First step: estimation of* ***single covariate models*** *and inclusion of all variables significant at the level of 0.05 in the full model*

- *Second step: verification of the model using the* ***deviance*** *and* ***Pearson statistics*** *which led to the decision to use correction for* ***overdispersion***

- *Third step: verification of the* ***linear relation between continuous covariates and logit*** *using the graphical method*

Calculations performed using SAS® Base 9.3

RESULTS AND INTERPRETATION

Table 1. Final model estimation

Parameter	DF	Estimate	Standard Error	Chi-Square	P-value
Intercept	1	-2.0739	0.3253	40.6418	<.0001
Age	1	0.00831	0.00444	3.5061	0.0611
Cigarettes use (years)	1	0.00999	0.0018	30.6668	<.0001
Alcohol use at least 3 times a week	1	0.1515	0.0692	4.7921	0.0286
Physical activity at least once a week	1	-2.0984	0.3943	28.3213	<.0001
Age*Physical activity	1	0.0229	0.00561	16.5892	<.0001

Source: Own analysis based on SHARE Wave 4 data

- *Each additional year of **smoking increases risk** of hospitalization by approximately **1%**,*

- *Using **alcohol** at least three times a week **increases this** risk by app. **16.5%** as compared with respondents who do not drink that often,*

- *Respondents who undertake **physical activities** at least once a week, assuming average age in the sample (i.e. 64.294 years), are approximately **50% less likely** to be taken into hospital, however as respondents get older, this difference diminishes.*

RESULTS AND INTERPRETATION

Table 1. Final model estimation

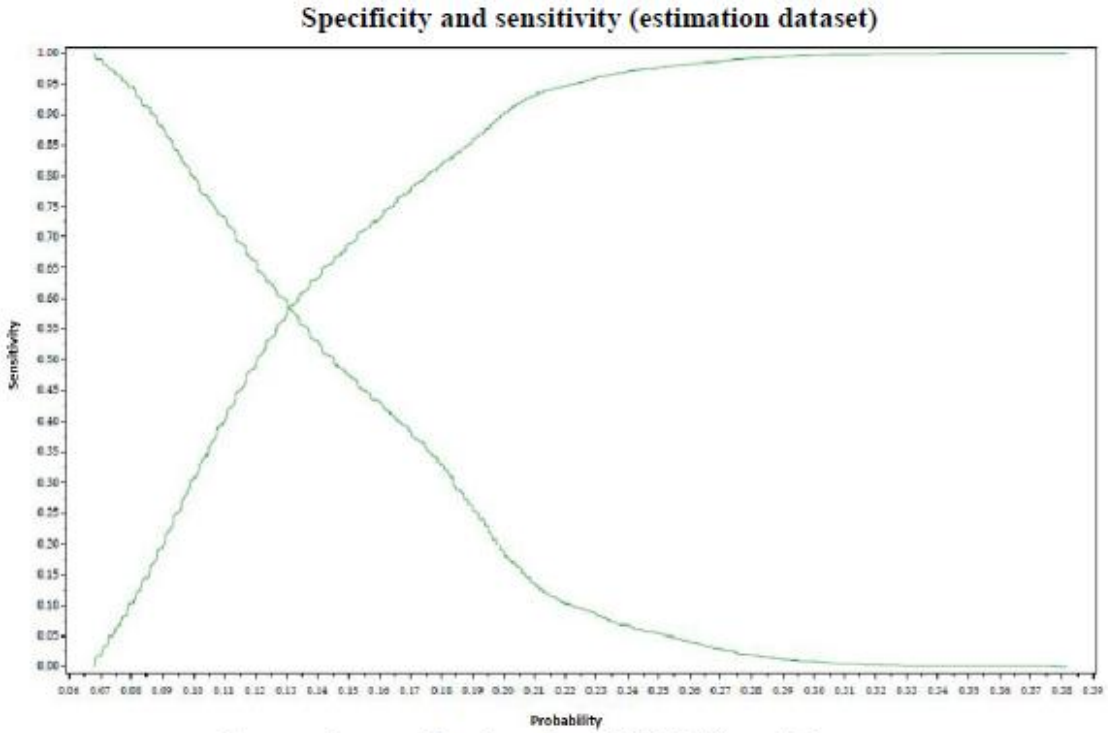
Parameter	DF	Estimate	Standard Error	Chi-Square	P-value
Intercept	1	-2.0739	0.3253	40.6418	<.0001
Age	1	0.00831	0.00444	3.5061	0.0611
Cigarettes use (years)	1	0.00999	0.0018	30.6668	<.0001
Alcohol use at least 3 times a week	1	0.1515	0.0692	4.7921	0.0286
Physical activity at least once a week	1	-2.0984	0.3943	28.3213	<.0001
Age*Physical activity	1	0.0229	0.00561	16.5892	<.0001

Source: Own analysis based on SHARE Wave 4 data

- *While comparing two respondents, out of whom one is 10 years older, the **older** respondent is app. **37% more likely** to be hospitalized **among respondents who undertake physical activities** at least once a week,*

- *Similar comparison among respondents who are not physically active gives result at the level of **9%**, thus it can be stated that **age influences risk of hospitalization to a larger extent among people who actively spend time** rather than among those who are not likely to perform activities requiring physical effort that often.*

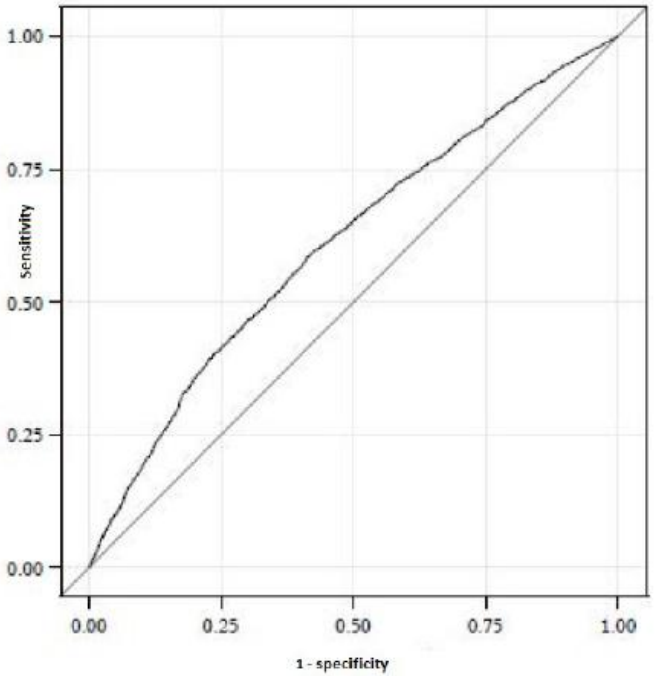
RESULTS AND INTERPRETATION



Source: Own analysis based on SHARE Wave 4 data

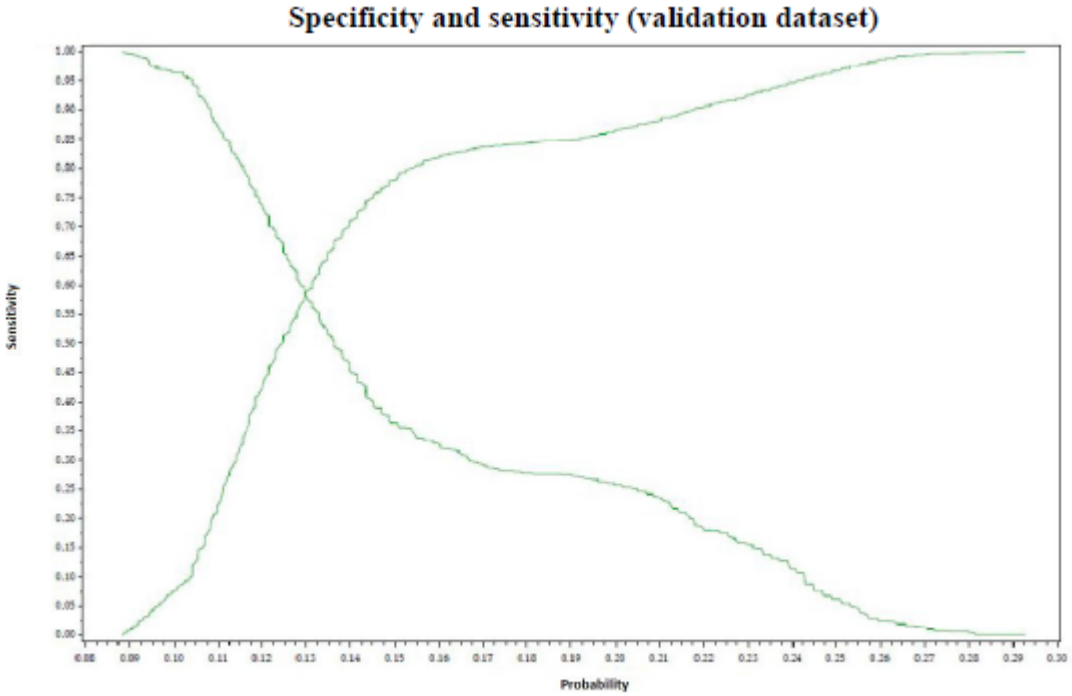
AUC calculated (0.6091 for the estimation dataset and 0.6181 for validation dataset) is not satisfying, however in this situation setting cut off point at relatively low value results in high percentages of true-positive, which is desired in this case

Figure 4. ROC curve (estimation dataset)



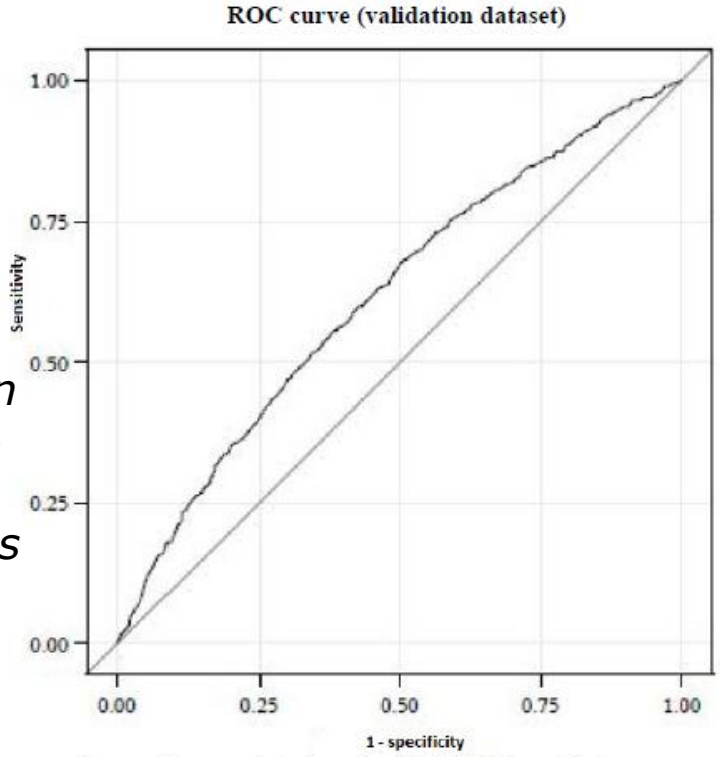
Source: Own analysis based on SHARE Wave 4 data

RESULTS AND INTERPRETATION



Source: Own analysis based on SHARE Wave 4 data

I.e. it is better to mark more respondents as being in risk of hospitalization and be sure that nearly 100% of those that in fact would be hospitalized will be in this group than to mark relatively few respondents as being in risk and lead to the situation where low percentage of respondents who in fact would be hospitalized is identified correctly).



Source: Own analysis based on SHARE Wave 4 data

CONCLUSIONS

Conducted analyses identified the following factors that have significant impact on the risk of hospitalization among people aged 50+ on the basis of SHARE Wave 4 data:

- **Smoking**
- **Drinking alcohol *(at least three times a week)***
- **Physical activities *(at least once a week)***
- **Age**
  - ***Age interacts significantly with physical activity =>***
    Effect of physical activity on the risk of hospitalization depends on age of the respondent

CONCLUSIONS

Model assessment:

- Final model included variables significant at the level of 0.05 (or their significant interactions)
- Correction for overdispersion was required
- Estimated logistic regression model is well-fitted to empirical data
  but…
- Model predictive power is not satisfying

Area for further research:

- Construction of a model with better predictive power

Thank you
for your attention!