JA PreventNCD
Joint Action Prevent Non-Communicable Diseases

Task 8.4 c: Modelling health care costs at micro-economic level

Cristina Mollica

Department of Statistical Sciences – Sapienza University of Rome

JA PreventNCD T8.3 & T8.4 Meeting

5th November 2024 – Istituto Superiore di Sanità, Rome

Co-funded by
The European Union

Outline

JA PreventNCD

- **Objectives** of cost data modelling
- Main **features of medical cost data**
- Review of the main **regression approaches** for medical cost data
- **Weighted gamma regression (WGR)**: motivations and advantages
- **Application** of WGR to colon cancer cost data from Italian registries
- **Conclusions**

Co-funded by
The European Union

Outcome of interest

JA PreventNCD

Outcome variable = individual annual cost of colon cancer patients

Our modelling analyses are distingued by:

- **the 3 annual phases of care (POCs)**:
  - **initial** (12 months from diagnosis)
  - **continuing** (6 months before to 6 months after the prevalence date)
  - **final** (12 months before death)
- **the 3 main healthcare services**:
  - **hospital admission (HA)**
  - **out-patient services (OPS)**
  - **territorial drug-prescriptions (DP)**

Co-funded by
The European Union

Cost drivers

JA PreventNCD

The main factors affecting medical expenses include

- **patient's socio-demographics**: *gender, age*
- **clinical conditions**: *disease subtype and stage, distance from diagnosis, comorbidities (Charlson index)*
- **treatment practices**: *number of treatments, treatment type*
- **geographical variation**: *cancer registry* (CR)
    - *treatment providers*
    - *local assistance administrations*
    - *regional insurance reimbursement rates*

Co-funded by
The European Union

Objectives

JA PreventNCD

A statistical modelling analysis of healthcare expenses data aims to

- **quantifying** the economic burden and **describe the variability** of cost data
- **identifying the determinants** of healthcare costs
- **making predictions** of cost impact under specific scenarios and interventions

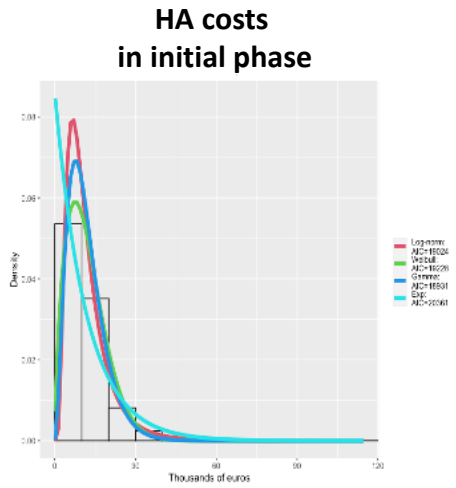
Co-funded by
The European Union

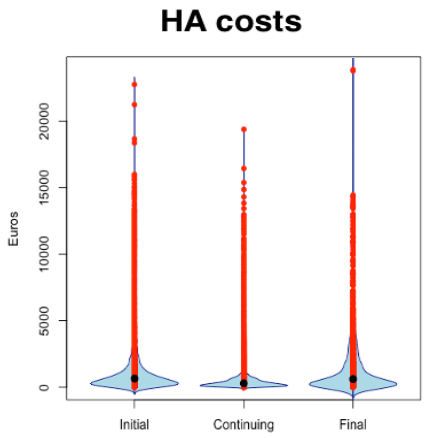
Features of medical cost data

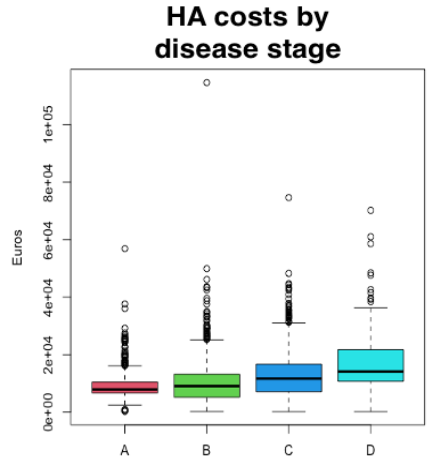
JA PreventNCD

Healthcare expense data are typically characterized by

- **right-skewness**: remarkable asymmetric distribution with a long heavy right tail
- **heteroscedasticity (non-constant variance)**: induced by one of more covariates
- **outliers**: patients with an atypically high amount of expenses



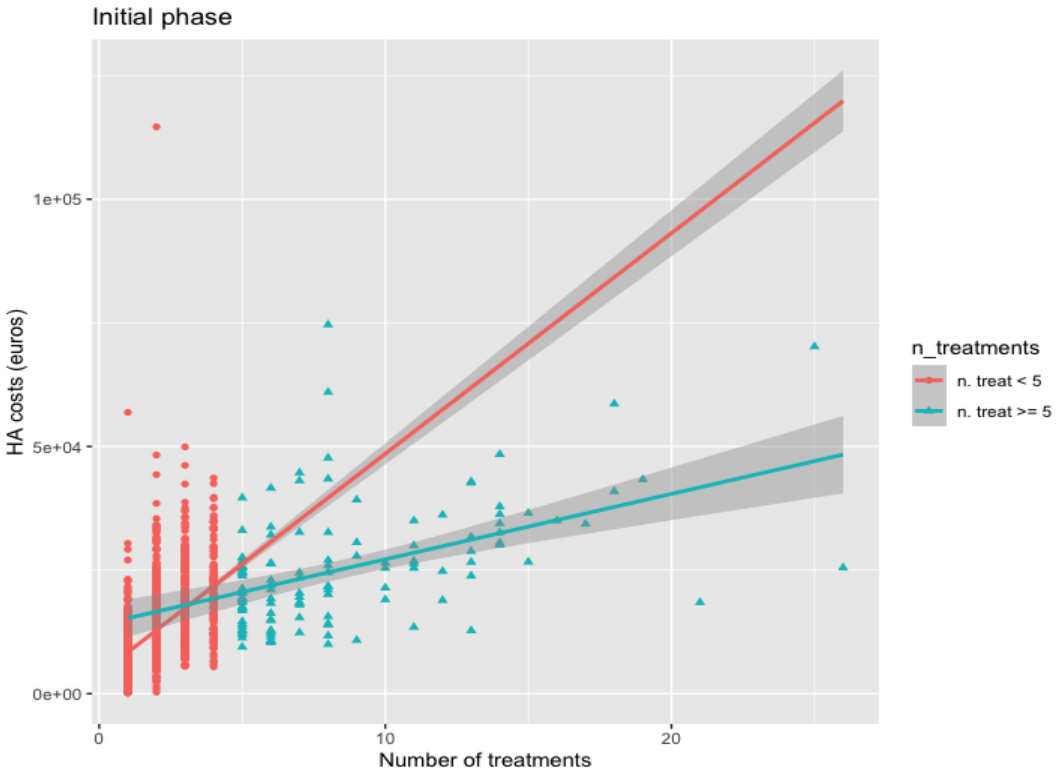


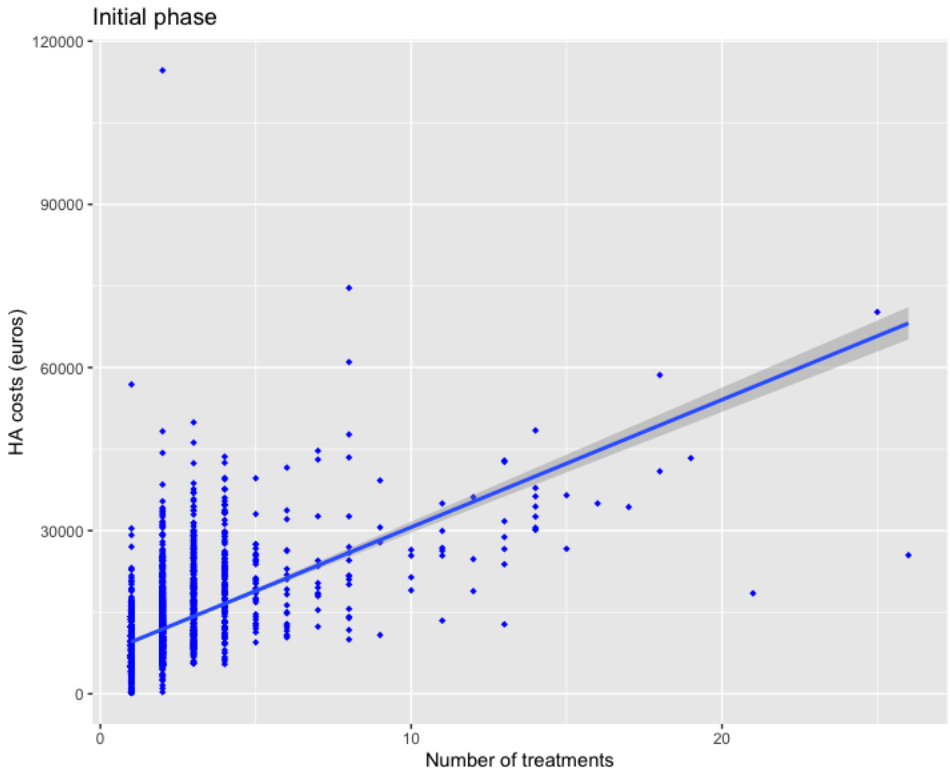


Co-funded by
The European Union

Features of medical cost data

JA PreventNCD





Co-funded by
The European Union

Statistical modelling issues

JA PreventNCD

Inadequacy of standard regression modelling approaches due to lack of robustness to

- **departures from the underlying assumptions**: biased and inconsistent estimates
- **presence of outliers**: maximum likelihood estimates are attracted by deviating observations

Need of suitable statistical methods to obtain reliable estimates!

Co-funded by
The European Union

Brief review of cost data models

JA PreventNCD

A variety of strategies have been proposed in the literature

- **data-transformation**: typically the log-transformation to mitigate skewness and heteroscedasticity
- **Generalized linear models (GLM)**: relaxation of the normality assumption in favour of the more general exponential family (for example, Gamma and Weibull)

Co-funded by
The European Union

Robust regression model

JA PreventNCD

We opt for a weighted regression model that

- **fits cost data on the original scale**
  - to gain a direct interpretability
  - to avoid the retransformation bias
- **accounts for heteroscedasticity**
- **incorporates the information provided by outliers** in a robust way

Co-funded by
The European Union

Weighted Gamma regression with log link

We propose to use the **weighted gamma regression (WGR)** with log link (Cantoni and Ronchetti, 2006)

- **right-skewness is captured** by the asymmetric shape of the Gamma density
- **heteroscedasticity is automatically addressed** by the mean-variance relationship of the Gamma density

$$E(Y_i \mid x_i) = \mu_i \qquad VAR(Y_i \mid x_i) = c\,\mu_i^2$$

where μ_i denotes the individual expected cost of patient *i*. The WGR, hence, postulates heteroscedasticity

$$\log(\mu_i) = \mathbf{x}'_i\boldsymbol{\beta} \implies VAR(Y_i \mid x_i) = c\exp(\mathbf{x}'_i\boldsymbol{\beta})$$

- **weighted least squares estimation procedure** $\Longrightarrow$ outliers are downweighted $(\omega_i < 1)$

Cantoni and Ronchetti (2006). A robust approach for skewed and heavy-tailed outcomes in the analysis of health care expenditures. Journal of Health Economics, 25(2), 198-213.

Co-funded by
The European Union

Weighted Gamma regression with log link

JA PreventNCD

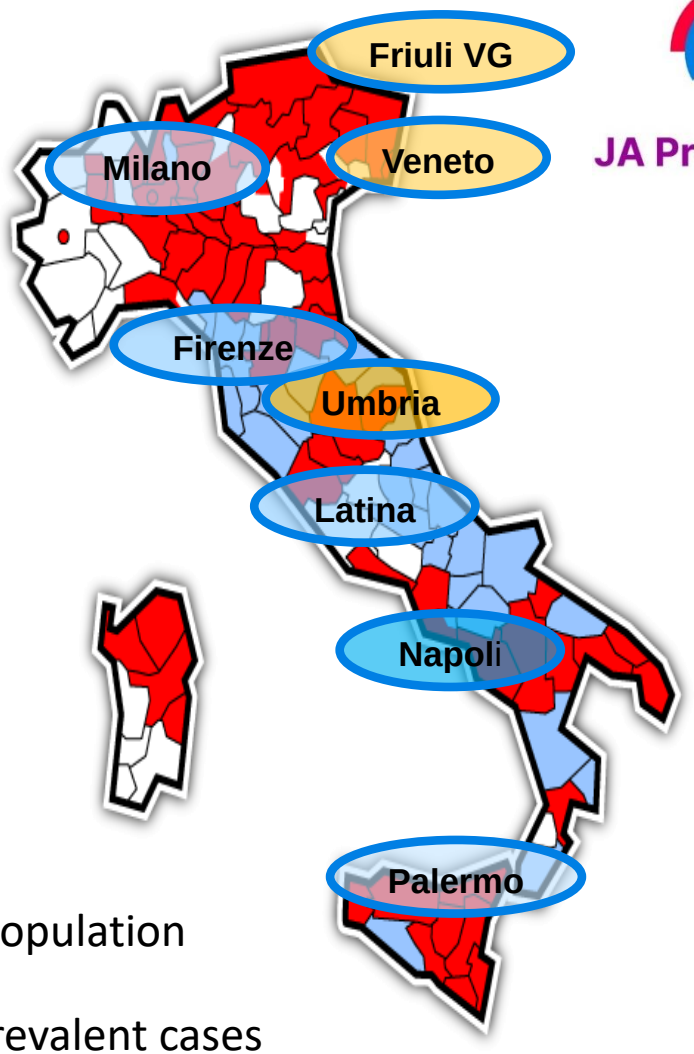
- **more flexible tool:** compromise between completely excluding outliers from the sample and including all the observations in the inferential process by assigning them the same importance
- **no artificial inflation of the regression coefficient SEs:** possibility to highlight significant effects which are lost with the classic Gamma regression
- the relative weights provide a **diagnostic tool for outlier detection**

Co-funded by
The European Union

Data sources

We analyze colon cancer data from **8 cancer registries** (CRs), namely

1. **Firenze**
2. **Friuli Venezia Giulia**
3. **Latina**
4. **Milano**
5. **Napoli**
6. **Palermo**
7. **Umbria**
8. **Verona**



JA PreventNCD

- covering over 10 million people, thus about one sixth of the Italian population
- coming from **EPICOST 1** project including a study cohort of 21,542 prevalent cases
- diagnosis of a malignant colon Cancer (ICD9-CM C18) in referred to the period 2010–2011

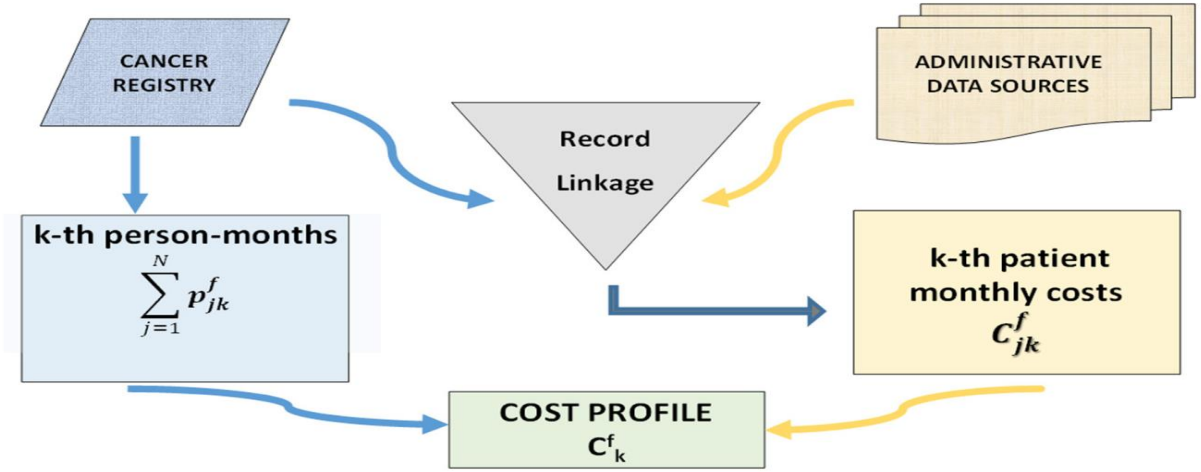
Co-funded by
The European Union

Data sources

JA PreventNCD

Information from the **CRs were linked with 3 administrative databases of healthcare services**:

- **Hospital Admission** (**HA**, or SDO in ita.)
- **Outpatient Services** (**OPS**, or SPA in ita.)
- **Drug Prescriptions** (**DP**, or FT in ita.)



Co-funded by
The European Union

Available information

JA PreventNCD

In order to prepare our data for the WGR on costs, we applied some **data cleaning**, **data trasformation** and **variable selection**:

- **patients' socio-demographics**: *gender* (M, F), *age at prevalence* (14-49, 50-74, 75-79, 80+)
- **clinical conditions**: *disease subtype* (proximal, distal), *comorbidities* (*Charlson index* = 0, 1, 2+, or not classified), *stage* (I, II, III, IV), distance from diagnosis (1-2, 2-3, 3-4, 4-8)
- **treatment practices**: *number of treatments*, *type of treatment* and *related costs*
- **geographical factor**: *cancer registry* (FI, FVG, LT, MI, NA, PA, UM, VE)

Co-funded by
The European Union

Sample sizes

JA PreventNCD

- data were **originally collected monthly** → Sum over the 12 months to compute **annual costs**
- **analysis by healthcare service type and POC**

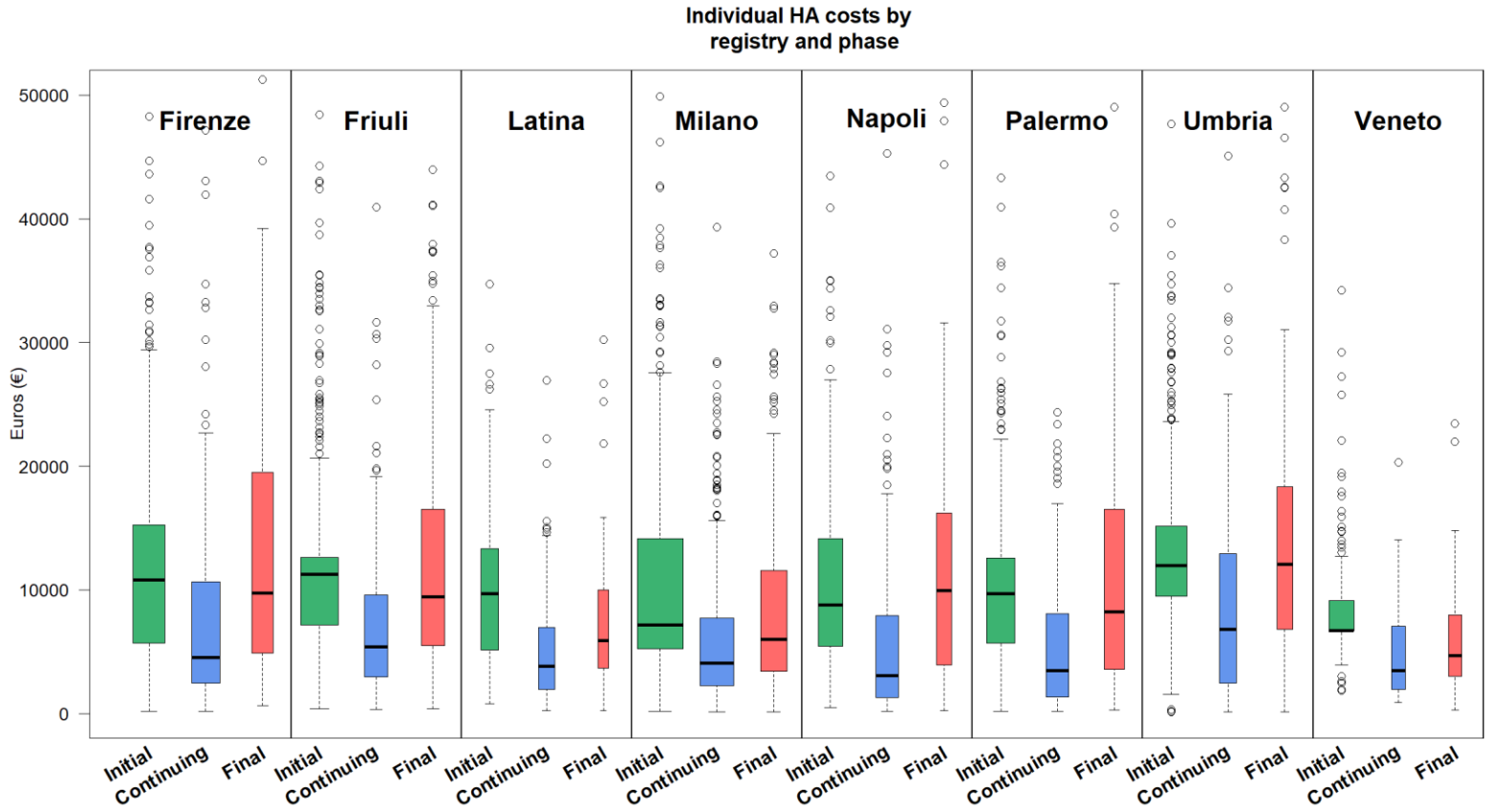
3 HEALTHCARE SERVICES * 3 POCs = 9 DATASETS

	initial	continuing	final
HA	2947	1739	1171
OPS	3085	14469	1598
DP	1164	3512	1186

Co-funded by
The European Union

Data exploration

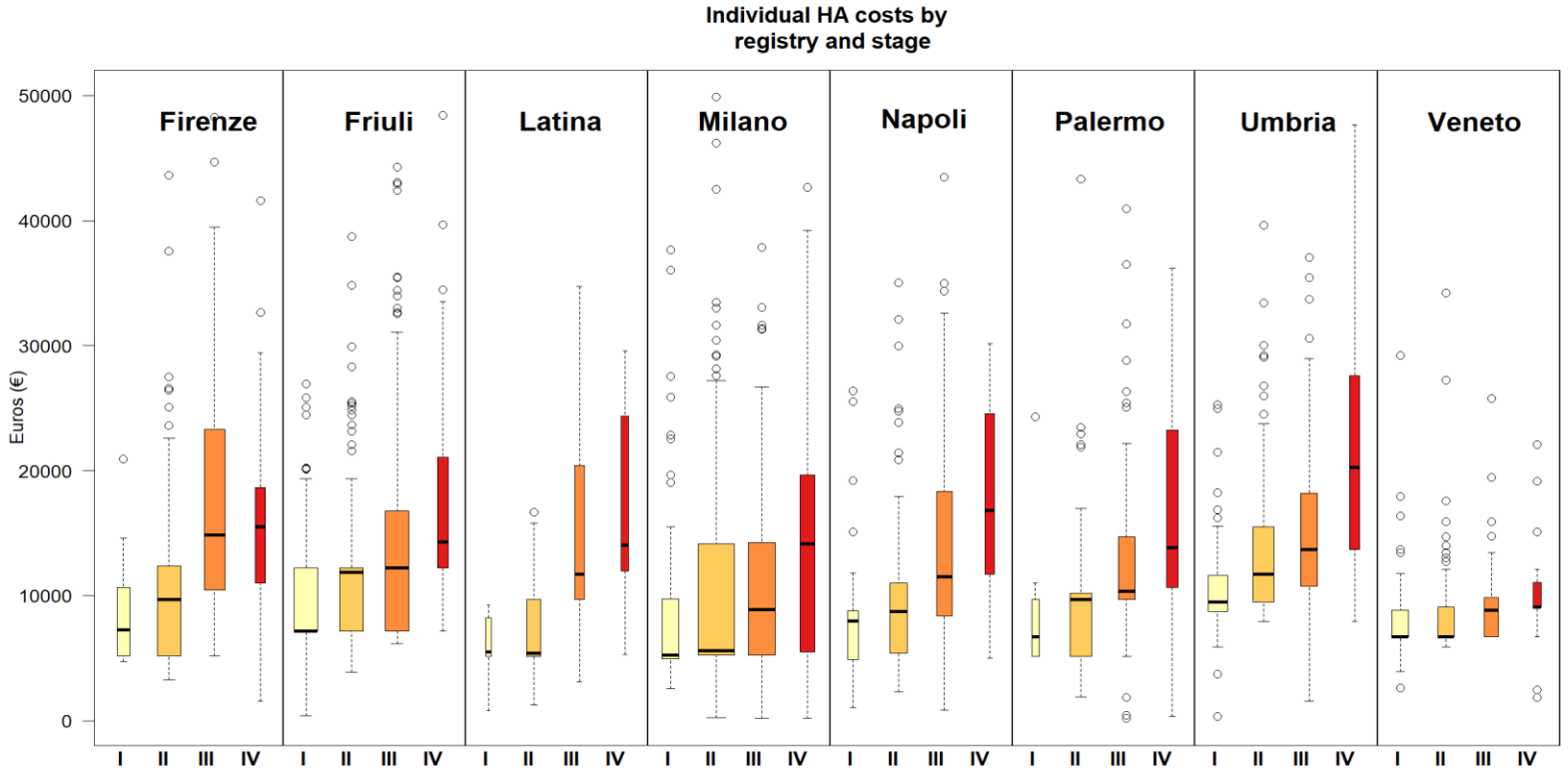
JA PreventNCD



Co-funded by
The European Union

Data exploration

JA PreventNCD



Co-funded by
The European Union

WGR model fitting

JA PreventNCD

We implement the WGR model in the **open-source R statistical environment**:

- use of the `glmrob` function from the contributed `robustbase` package
- example of R code for HA data in initial phase

```
mod_pesi2_sdo_iniz = glmrob(costs_anno ~ sesso + eta_prev_agg1 + registro + stadio_agg + sotto_sede +
Charlson_3lv + n_tratt_anno_centr + n_tratt_anno_centr*registro, family = Gamma(link = "log"), data =
dati_sdo_anno_iniz, method = "Mqle", weights.on.x = "hat", control = glmrobMqle.control(tcc= 1.5))
```

- *method = "Mqle"* is the Huber type robust estimator
- *weights.on.x = "hat"* is the robust function specification with $w(x_i) = \sqrt{1 - H_{ii}}$ (h_{ii} element of diagonal matrix)
- *control = glmrobMqle.control(tcc=1.5)* is a tuning to handle the iterative fitting estimation process

see Cantoni and Rocchetti 2006 and R documentation for more details.

Co-funded by
The European Union

Estimated effects

JA PreventNCD

- num. of treatments and disease stage have **always a significant positive effect** on average costs: higher num. of treatments or cancer stage increase expenses.
- age at prevalence has a **significant negative effect**: older patients are less costly.
- distance from diagnosis is similar to age at prevalence, but with a **significant negative effect** only when several years have passed from diagnosis.
- comorbidities have a **significant positive effect** on HA costs in initial phase, but a **significant negative effect** on OPS and DP costs.
- cancer subsite is **not significant.**
- cancer registries have **always significant effects**, with an associations varying for the different cost components consistently with the characteristics of the local healthcare administrations.

Co-funded by
The European Union

Main differences with classic Gamma regression

JA PreventNCD

- **cost drivers identification**: the WGR highlighted sex as a significant predictor of OPS costs for all POCs, with females having significantly lower expenses than males, whereas the estimated effect is not significant with the GR.

- **improvement of goodness-of-fit**:

BIC values are better with the WGR than the classic GR in 7 out of 9 cases

		WGR	GR
HA	initial	**57726**	58225
	continuing	**32707**	33767
	final	**22706**	23742
OPS	initial	**49624**	50421
	continuing	202913	**202017**
	final	**24865**	25740
DP	initial	**11734**	12212
	continuing	33688	**32666**
	final	**13954**	14536

Co-funded by
The European Union

Conclusions

JA PreventNCD

Final remarks:

- **usefulness of regression modelling** for evaluating cancer cost levels and drivers
- description of **robust estimation of Gamma regression**
- several **theoretical and practical advantages**
- **better detection of influential factors**, including meaningful territorial patterns

Co-funded by
The European Union

Future research

JA PreventNCD

Possible **future developments**:

- application to **other cancer types**
- **additional covariates**: more specific categorization of the cancer type
- use of the weights to **characterize patients with outlying costs**
- **other robust strategies**: modelling median instead of mean costs
- development of an **R package** for modelling cost data at micro-level

Co-funded by
The European Union

References

1. Cantoni E. and Ronchetti E. (2001). *Robust inference for generalized linear models*. Journal of the American Statistical Association 96, 1022–1030.
2. Cantoni, E. and Ronchetti E. (2006). *A robust approach for skewed and heavy-tailed outcomes in the analysis of health care expenditures.* Journal of Health Economics, 25(2), 198-213.
3. Duan N., Manning W. G., Morris C. N. and Newhouse, J. P. (1983). *A comparison of alternative models for the demand for medical care*. Journal of Business & Economic Statistics, 1(2), 115-126.
4. Gilleskie D. B. and Mroz T. A. (2004). *A flexible approach for estimating the effects of covariates on health expenditures*. Journal of health economics, 23(2), 391-418.
5. Manning W. G. and Mullahy J. (2001). *Estimating log models: to transform or not to transform?* Journal of health economics, 20(4), 461-494.
6. Mihaylova B. et al. (2011). *Rewiew of statistical methods for analysing healthcare resources and costs.* Health Economics, 20(8), 897-916.
7. Petrinco M. et al. (2012). *Robust gamma regression models for the analysis of health care cost data.* Model Assisted Statistics and Applications, 7(2), 115-124.
8. Zhou Q. M. and Song T. (2015). *Profiling heteroscedasticity in linear regression models.* Canadian Journal of Statistics, 43(3), 358-377.

Co-funded by
The European Union

JA PreventNCD

Thank you!

cristina.mollica@uniroma1.it

This document is part of a project that has received funding from the European Union's EU4Health programme under grant agreement No 101128032. The information reflects only the authors' view and the European Commission is not responsible for any use that may be made of the information it contains.