Italy COVID-19 analysis of vaccine coverage reasons
LUMSA University

Matteo Fasulo*
Tecniche Informatiche per la gestione dei Dati
LUMSA University, Rome, Italy
Email: *m.fasulo1@lumsastud.it

***Index Terms*—COVID-19 Vaccines, Socioeconomic factors, Italy, Vaccination coverage**

I. Introduction

The COVID-19 pandemic has highlighted existing health inequalities of those living in more socioeconomically deprived areas in Italy. Since there are severe outcomes in these groups, equitable vaccination coverage should be prioritised. The aim of this research was to identify inequalities in coverage of COVID-19 vaccination in Italy and to highlight areas which may benefit from targeted measures. Records about vaccines administrations from Presidenza del Consiglio dei Ministri were linked to Social Vulnerable Index (SVI) [1] of ISTAT. SVI was derived from the 2011 census while GDP from the 2020 report. Administrations of first dose of any COVID-19 vaccine was analysed over time by age group, region of residence, GDP and SVI divided into seven distinct factors. While linear regression models were adjusted for age group, trying to predict the coverage for each region, logistic regression models were used to highlight the feature importance. This study included all the vaccine administrations since 27th December 2020. The vaccination for kids aged between 5 and 12 starts from 15th December 2021. Mean uptake of second dose of COVID-19 vaccine among regions was $\simeq 74\%$. After adjustment the odds of being vaccinated were lower for individuals who live in regions with high levels of house crowding, youth unemployment and medium-low education level. The largest inequality was observed among regions with greater economic difficulties, greater social hardship and a lower average level of education than those without these characteristics.

II. Dataset

The original dataset is available at GitHub [2] while GDP and SVI at ISTAT website [3]. The information about uptake of first and second doses are referred to the period between 27th December 2020 and 25th January 2022. It contains more than 175000 rows each regarding the number of administered doses, for a specific age-group, in a region on a day by vaccine manufacturer. There are 14 attributes in this dataset:

- data_somministrazione - The date of vaccines administration
- fornitore - The manufacturer of vaccine
- area - The region in which doses has been administered
- fascia_anagrafica - The age group of individuals
- sesso_maschile - Total number of male individuals who has been vaccinated during that day
- sesso_femminile - Total number of female individuals who has been vaccinated during that day
- prima_dose - Total number of first doses administered during that day
- seconda_dose - Total number of second doses administered during that day
- pregressa_infezione - Total number of individuals who had already had the disease before getting vaccinated
- dose_addizionale_booster - Total number of third doses administered during that day
- codice_NUTS1 - Specific code for region
- codice_NUTS2 - Specific code for region
- codice_regione_ISTAT - Specific code for region by ISTAT
- nome_area - The full name of region in which doses has been administered

Since some of these characteristics cannot be used for the purpose of the research, they will be dropped in order to have only information about second doses (vaccinated individuals). The area variable (also codice_NUTS1 and codice_NUTS2) are redundant and the gender is referred to all the doses administered in one day by a specific region; also pregressa_infezione, prima_dose and dose_addizionale_booster are unnecessary here. The final dataset will contains any features that have not been discarded from the original dataset plus the GDP of regions with their residents and the 7 factors SVI [1]:

- Single-parent families: the overall incidence of single parents household (less than 65 years old).
- Families with at least 6 members: the percentage incidence of families with six or more members.
- Population without university-type education: people between 25 and 64 years old who have completed at least the upper secondary school.
- Population with welfare problems: families with at least two members (aged 65 or more) having a member aged

80 or more.
- Overcrowded population: families having a dwelling floor space of 40 sqm (Square Meters) [1] made of at least four members; dwelling floor space between 40 and 59 sqm with five members or more; dwelling floor space between 60 and 79 sqm with six members or more.
- Population under 30 who do not study or work: people aged between 15 and 29 who are not working, not looking for a job and not attending any training course (NEET (Neither in employment nor in Education and Training) [2]).
- Population with economic hardship: families with young sons and adults (aged less than 64) where nobody is employed or is receiving a pension.

For the descriptive statistics of the selected numeric features (see Figures VII).

III. Libraries

The following section will itemise all the libraries used for this research as well as their programming languages that implement them. Each of the libraries will redirect to the maintainer website for a more precise description of the characteristics.

A. Python

- Pandas - Manipulation of dataset and their transformations.
- Numpy - Manipulation of datatypes as well as data formatting for typos.
- Seaborn - Data visualization library based on matplotlib for quick render graphs.
- Statsmodels - Python module that provides classes and functions for the estimation of many different statistical models.
- Matplotlib - Library for creating static, animated, and interactive visualizations in Python.
- Json - Data interchange format inspired by JavaScript object literal syntax.
- Os - Module that provides a portable way of using operating system dependent functionality.
- Urllib - Module that defines functions and classes which help in opening URLs (mostly HTTP) in a complex world — basic and digest authentication, redirections, cookies and more.

B. R

- GGPlot2 - A system for declaratively creating graphics, based on The Grammar of Graphics book.
- Dplyr - A grammar of data manipulation, providing a consistent set of verbs that help you solve the most common data manipulation challenges.
- Magrittr - The magrittr package offers a set of operators which make your code more readable.
- Plotly - Plotly's R graphing library makes interactive, publication-quality graphs.
- Jsonlite - A tool for reading json objects in R.
- Gamlss - Generalized Additive Models for Location, Scale and Shape.
- Stargazer - An R package that creates LATEX code, HTML code and ASCII text for well-formatted regression tables, with multiple models side-by-side, as well as for summary statistics tables, data-frames, vectors and matrices.

[1] Square Metre - https://en.wikipedia.org/wiki/Square_metre
[2] NEET - https://en.wikipedia.org/wiki/NEET

IV. Method

In order to get the importance of each feature in the dataset, both linear and logistic regression have been used but while linear regression tries to predict the actual coverage value as a function of GDP and SVI indexes, the logistic one models a binary variable of region's coverage above or below average coverage. Indeed all the regions who have a coverage above the average will be indicated with 1 while the others with 0. A stepwise approach has been used for feature selection preferring the model with the highest R^2.

A. Linear Regression

In statistics, linear regression is a linear approach for modelling the relationship between a scalar response and one or more explanatory variables (also known as dependent and independent variables). The relationships are modeled using linear predictor functions whose unknown model parameters are estimated from the data. Such models are called linear models.
Given a data set $\{y_i, x_{i1}, \ldots, x_{ip}\}_{i=1}^{n}$ of n statistical units, a linear regression model assumes that the relationship between the dependent variable y and the p-vector of regressors x is linear. This relationship is modeled through a disturbance term or error variable ϵ — an unobserved random variable that adds "noise" to the linear relationship between the dependent variable and regressors [4]. Thus the model takes the form:

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \epsilon_i, \text{where } i = 1, \ldots, n \quad (1)$$

In this specific case, the linear equation for coverage will be:

$$\text{coverage}_i = \beta_0 + \beta_1 \text{GDP}_{i1} + \ldots + \beta_p \text{eco\_Unease}_{ip} + \epsilon_i,$$

$$\text{with } \; i = 1,\ldots,n \quad (2)$$

B. Logistic Regression

Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist. Mathematically, a binary logistic model has a dependent variable with two possible values, such as pass/fail which is represented by an indicator variable, where the two values are labeled "0" and "1". In the logistic model, the log-odds (the logarithm of the odds) for the value labeled "1" is a linear combination of one or more independent variables ("predictors"); the independent

variables can each be a binary variable (two classes, coded by an indicator variable) or a continuous variable (any real value) [5]. Recalling that in linear models:

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \epsilon_i, \text{where } i = 1, \ldots, n \quad (3)$$

For classification, we prefer probabilities between 0 and 1, so we wrap the right side of the equation into the logistic function. This forces the output to assume only values between 0 and 1 [6].

$$P(y_i = 1) = \frac{1}{1 + \exp\left(-(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip})\right)} \quad (4)$$

V. Results

The first part of the results regard the proportions between vaccine manufacturer as well as the most manufacturer per region and both per region and age-group. For the first doses (see figure 1), the most common manufacturer was Pfizer/BioNTech ($\sim$ 68%) followed by Moderna ($\sim$ 14%) and Vaxzevria "AstraZeneca" ($\sim$ 13%); Janssen (Johnson & Johnson $\sim$ 3%) and Pfizer Pediatrico (age-group 5-11 years $\sim$ 2%).
For the second doses (see figure 2), the most common manufacturer was Pfizer/BioNTech ($\sim$ 72%) followed by Moderna ($\sim$ 14%) and Vaxzevria "AstraZeneca" ($\sim$ 13%); Janssen is missing since it was a one-dose vaccine and Pfizer Pediatrico (age-group 5-11 years $\sim$ 1%).
The regional effect has been inspected by aggregating the number of vaccines by each manufacturer per area. Trivially the most common manufacturer was Pfizer/BioNTech in all the areas (since it was the most used overall) but regarding the second most common manufacturer, there are regions who preferred Moderna over Vaxzevria and vice versa (see figure 3).
The age group effect has been inspected by aggregating the number of vaccines by each manufacturer per age-group. Once again, Pfizer/BioNTech was the most used (except for the age-group of 5-11 obviously) while Vaxzevria shows a growing trend from 20-29 to 70-79. Moderna was preferred by individuals under the age of 59 while Vaxzevria overcomes Moderna in older people (see figure 4). This could be caused by the vaccination campaign implemented by the Italian government suggesting that older people should be vaccinated with Vaxzevria. An interesting anaylis has been made with both age-group and region effect in the same graph (see figure 5).

The linear regression model was

$$\text{coverage}_i = \beta_0 + \beta_1 \text{GDP}_{i1} + \ldots + \beta_p \text{eco\_Unease}_{ip} + \epsilon_i,$$

with i = 1,...,n (5)

Regarding the coefficients:

Table I
Coefficients and CI for Linear Model

	Coef	2.5 %	97.5 %
(Intercept)	0.656	0.654	0.657
PIL	0.00000	0.00000	0.00000
mono_Family	0.005	0.005	0.005
six_more_Family	−0.015	−0.015	−0.014
low_Educ	−0.024	−0.024	−0.023
house_Crow	−0.023	−0.023	−0.023
ass_Unease	0.043	0.043	0.044
u30_Unemployed	−0.010	−0.010	−0.010
eco_Unease	0.040	0.040	0.041

The three most important factors that model coverage are:

1) Ass. Unease.
2) Eco. Unease.
3) Low Educ.

Considering the fact that the sign of a regression coefficient tells whether there is a positive or negative correlation between each independent variable and the dependent variable, the model assumes that factors like Mono Family, Ass. Unease and Eco. Unease tend to increase the coverage while Six More Family, Low Educ, House Crow and U30 Unemployed tend to decrease it. GDP was not considered since its effect can be considered around zero ($1.896 \cdot 10^{-7}$). This means that regions that have vaccinated the most are those with:

- High values of both Eco. Unease and Ass. Unease.
- More Mono Family than average.
- A low level of U30 Unemployed (more employed) and Low Educ (meaning that people are generally more educated).
- Low values of House Crow.

The stepwise approach (both in Python and R) selected the model containing all the feature described above with an R^2 of ~ 0.60.

The same considerations can be made for logistic model:

Table II
ODDS RATIO AND CI FOR LOGISTIC MODEL

	OR	2.5 %	97.5 %
(Intercept)	0	0	0
PIL	1.000	1.000	1.000
mono_Family	1.438	1.370	1.510
six_more_Family	0.0005	0.0004	0.0005
low_Educ	0.001	0.001	0.001
house_Crow	0	0	0
ass_Unease	51,976,584.000	48,050,553.000	56,223,396.000
u30_Unemployed	2.511	2.303	2.737
eco_Unease	4,572.147	3,877.677	5,390.992

As for linear model, the three most important factors are Ass. Unease, Eco. Unease and Low Educ. The interpretation of "mono_Family" factor for example is that for a one unit increase in mono_Family, the odds of being above the average coverage (versus being below) increase by a factor of 1.438.

VI. CONCLUSION

This analysis highlights the need of further researches about inequality between regions in Italy. In fact southern Italy revealed a general pattern of higher level of inequalities. Due to the fact that the seven factors of SVI were taken from the 2011 census, a more updated report from ISTAT would be preferred as the actual census might be different and by that, it could change the entire perspective of the research. In order to obtain a better accuracy in the development of the proposed models, it would be desirable to use regressors with greater granularity to capture the variations related to the coverage of each region. Furthermore, with a more in-depth analysis, it would be possible to outline information campaigns aimed at specific regions (or even more little areas) that could have more significant effects among the population about choosing the vaccination or not.

REFERENCES

[1] "Social Vulnerable Index of ISTAT," 2020. [Online]. Available: https://www.istat.it/it/files//2020/12/Le-misure-della-vulnerabilita.pdf
[2] "Covid19 OpenData Vaccini." [Online]. Available: https://github.com/italia/covid19-opendata-vaccini
[3] "Prodotto interno lordo lato produzione - ISTAT." [Online]. Available: http://dati.istat.it/Index.aspx?DataSetCode=DCCN_PILT
[4] "Linear regression - Wikipedia." [Online]. Available: https://en.wikipedia.org/wiki/Linear_regression
[5] "Logistic regression - Wikipedia." [Online]. Available: https://en.wikipedia.org/wiki/Logistic_regression
[6] "Logistic Regression - Interpretable Machine Learning," 2022. [Online]. Available: https://christophm.github.io/interpretable-ml-book/logistic.html

VII. Tables and Figures

	residenti	coverage	PIL	mono_Family	six_more_Family	low_Educ	ass_Unease	house_Crow	u30_Unemployed	eco_Unease
mean	3152579.95	0.74	88615.23	6.85	1.41	1.54	3.13	1.30	11.11	2.51
std	2461658.34	0.02	85586.31	0.77	0.47	0.73	0.49	0.63	4.13	1.90
min	124089.00	0.70	4522.40	5.50	0.70	0.70	2.20	0.60	6.70	0.90
25%	1201510.00	0.72	30759.10	6.40	1.10	1.10	2.70	1.00	8.10	1.10
50%	1860601.00	0.74	46194.70	6.80	1.40	1.30	3.10	1.10	9.20	1.40
75%	4833705.00	0.76	126374.60	7.10	1.60	1.60	3.60	1.50	12.20	3.30
max	9981554.00	0.78	367167.20	9.10	2.80	3.10	3.90	3.50	20.40	7.60

Figure 1. Pie Chart of Vaccine Manufacturer (1st dose)

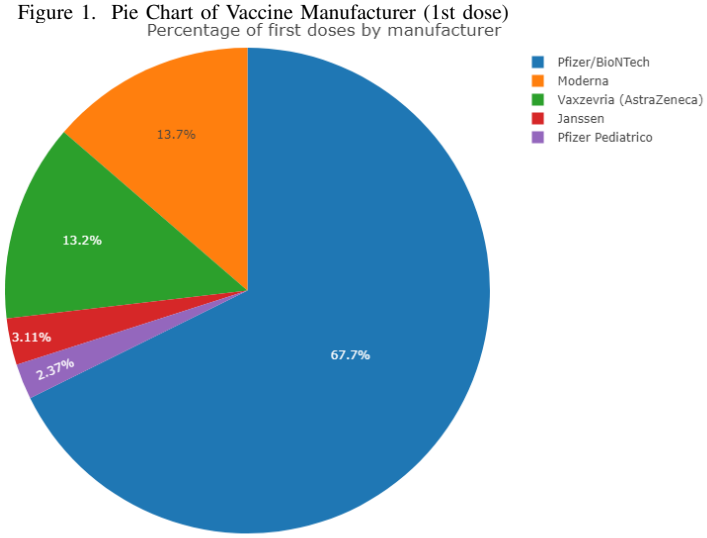


Figure 2. Pie Chart of Vaccine Manufacturer (2nd dose)

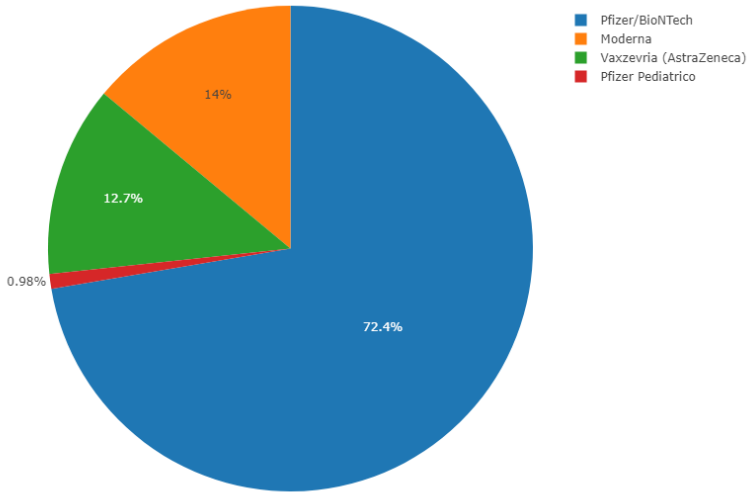


Figure 3.

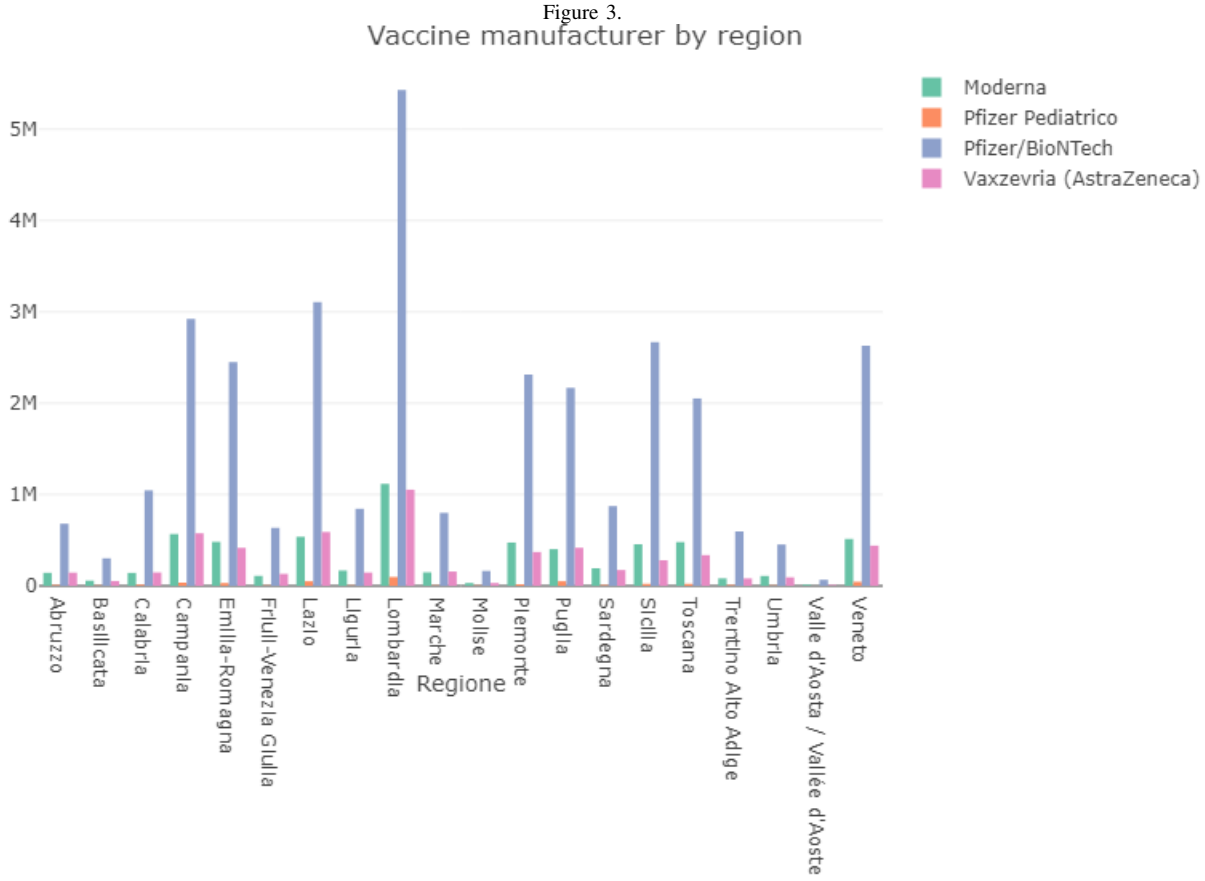


Figure 4.

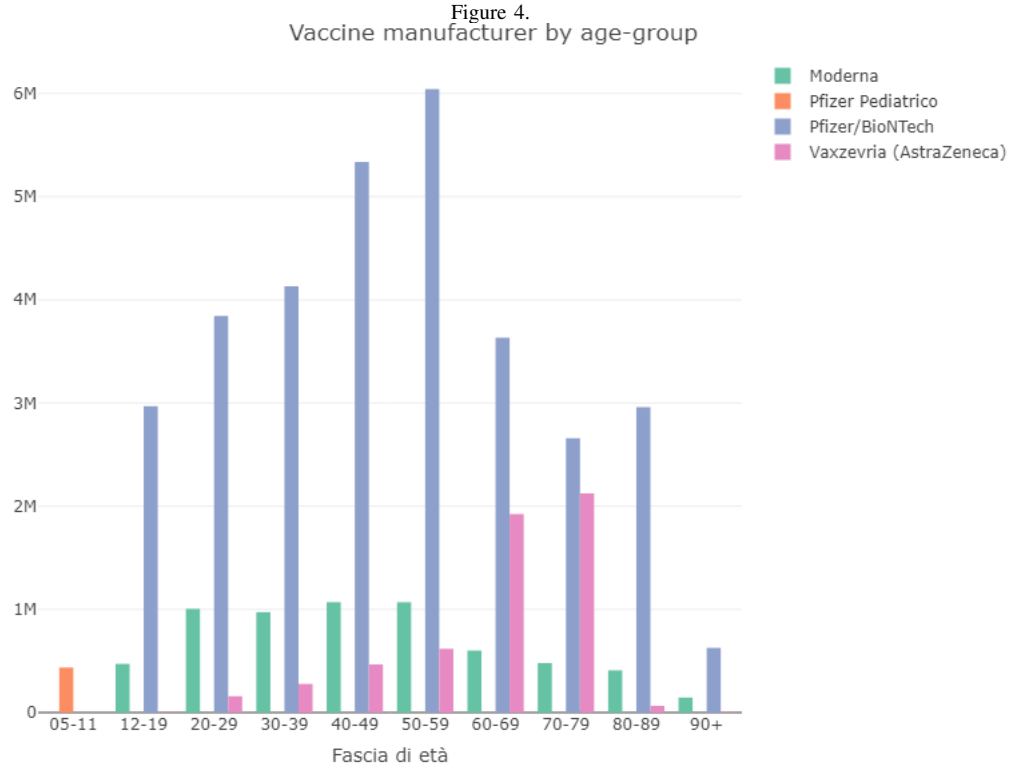
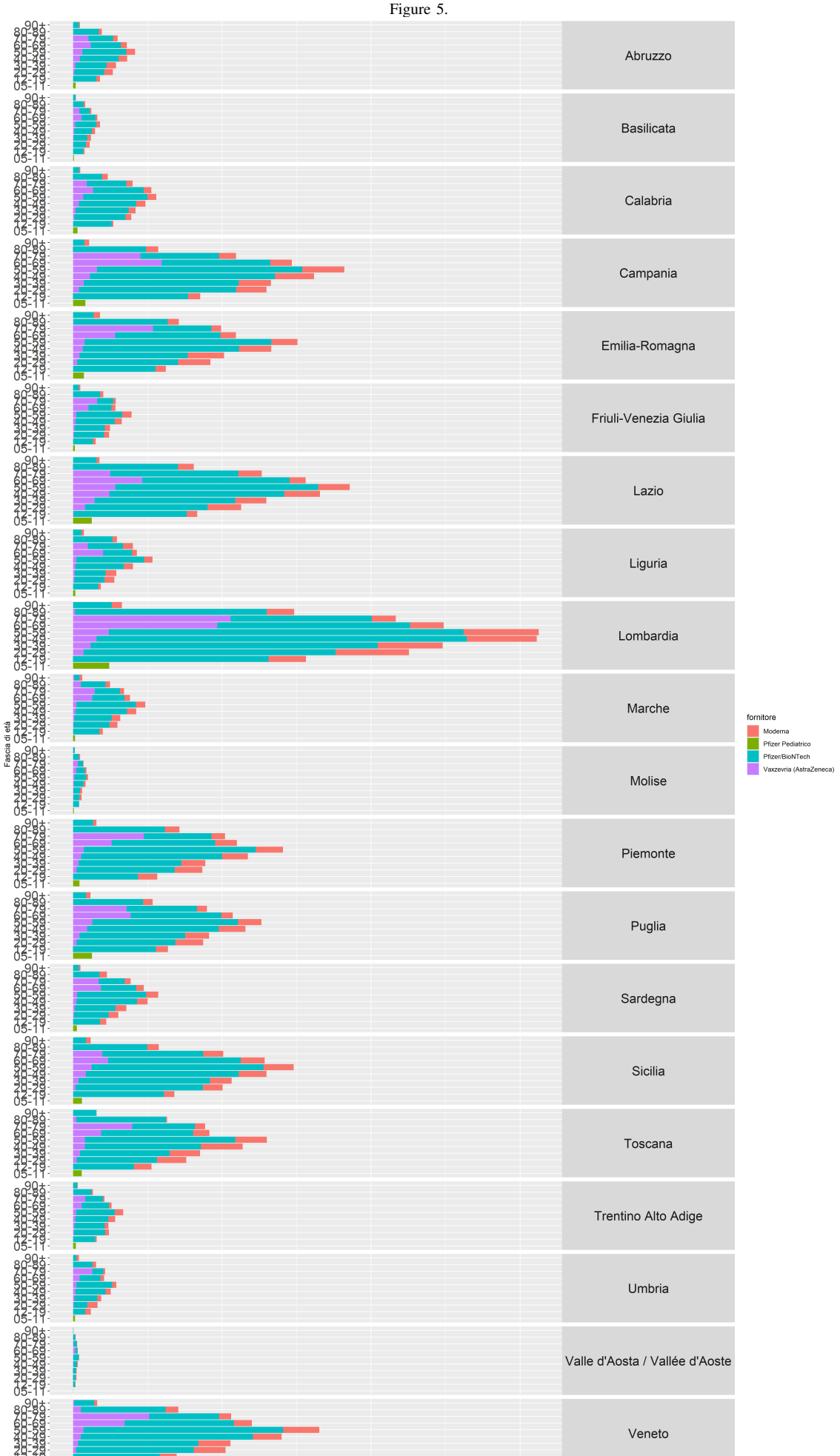


Figure 5.


Abruzzo
Basilicata
Calabria
Campania
Emilia-Romagna
Friuli-Venezia Giulia
Lazio
Liguria
Lombardia
Marche
Molise
Piemonte
Puglia
Sardegna
Sicilia
Toscana
Trentino Alto Adige
Umbria
Valle d'Aosta / Vallée d'Aoste
Veneto
Fascia di età
fornitore
Moderna
Pfizer Pediatrico
Pfizer/BioNTech
Vaxzevria (AstraZeneca)