Federal State Autonomous Educational Institution for Higher Education
«National Research University
«Higher School Of Economics»

Econometrics project

“Analysis of the influence of HIV treatment intensity on the share of patients with undetectable viral load using OLS regression ”

Prepared by the team №21:
Arzamastseva Mishel
Kononov Anatolii
Doronin Nikita
Levashov Danil
Murashko Lev

Saint Petersburg
2024

Table of content

Table of content........ 2

Abstract........3

Introduction........4

Literature overview........5

Data usage and methodology........ 7

ETL process, description, EDA and IDA........8

ETL process........8

Description of initial data........ 8

Transformation of data........9

Problem of missing values for certain years........ 13

Explanatory data analysis........14

Descriptive statistic........14

Visualization........ 15

Intelligence data analysis........20

Correlation matrix........20

Formulation and justification of the model........22

Regression analysis........ 24

Baseline model........ 24

Regression with control variables........ 25

Regression with individual (state) fixed effects........27

Analysis of results........28

Fixed effect of region........ 28

Answer to research question........ 29

Limitations of the study........30

Possible expansion of the study........32

Cross effect investigation........ 32

Assessing team members' contributions to group work........34

References........ 36

Abstract

This research analyzes the influence of HIV treatment intensity on the share of patients with undetectable viral loads in Russia, using Ordinary Least Squares (OLS) regression techniques. With over one million individuals affected by HIV in Russia, assessing the impact of treatment factors is crucial. The central hypothesis suggests that increased access to antiretroviral therapy (ART) and effective healthcare from infectious disease doctors leads to higher viral suppression rates.

The study utilizes data from the "Если быть точным" database, conducting an explanatory data analysis (EDA) to examine relationships between treatment variables and viral loads. Key variables include the proportion of patients on ART and the availability of infectious disease doctors.

Results indicate a significant positive correlation between the percentage of patients receiving ART and the share with undetectable viral loads, while the impact of infectious disease doctors was not significant when analyzed independently.

These findings highlight the importance of enhancing ART accessibility to improve health outcomes among HIV-positive individuals in Russia and support policies aimed at strengthening healthcare infrastructure to combat the HIV epidemic effectively.

Introduction

Russia is home to over one million people infected with the human immunodeficiency virus (HIV), hence we can say that it is a pandemic. However, with the appropriate medical intervention, HIV can be effectively managed, allowing patients to lead productive lives with minimal restrictions. Taking into account the fact that in the neglected cases HIV might lead to severe deaths, the issue of proper treatment is very significant in today’s Russia. Hence, in our work we are investigating how the availability of HIV treatment affects the number of people who have undetectable viral load in their blood (suppressed HIV).

The main hypothesis of our research is that the availability rate of ART courses and accessibility of care of infectious disease doctors together with stable drug supply have a significant causal relationship with the percentage of “recovered” patients.

We assume that mentioned variables have a positive relationship with the percentage of patients with undetectable viral load from all HIV+ patients. The reason is that people living with HIV receive most of their care directly from infectious disease doctors, who are the main providers of antiretroviral therapy. Moreover, the degree of permanent availability of ART courses is also thought to be an important positive influencer.

As a result, our research is expected to clarify what factors increase the number of people who suppressed HIV. Further research is needed to estimate the degree of influence of each key treatment factor and the power of casual relationship.

Literature overview

HIV, despite its first reported case being over 40 years ago in 1981, still remains one of the biggest healthcare issues to this day. According to the World Health Organization, there were approximately 40 million people living with HIV at the end of 2023 and more than 600 thousand people died of HIV-related illnesses in the same year.

However, there has been tremendous progress related to treating HIV and AIDS over the past 40 years. Nowadays antiretroviral therapy is at the focal point of this issue. ART has transformed a consistently fatal into a potentially chronic disease *(Volberding & Deeks, 2010).* More than 40 drugs were developed that aid in controlling infection in 2016, and as of 2021 the U.S. FDA approved approximately 222 antiretroviral drugs for AIDS treatment globally. Thanks to advances in ART, there are decreases in HIV infections in almost every geographic region *(Haris & Abbas, 2024)*. Different works show that immediate ART is associated with a 63% reduction in overall mortality among people living with HIV with CD4 counts >500 cells/µL. These studies show the significant negative impact of delays in ART initiation in a real-world setting *(Zhao et al., 2017)*. Additionally, more than 65% infectious disease doctors routinely treat people living with HIV in an outpatient setting and majority of these ID doctors acted as PWLH's primary care physicians *(Lakshmi et al., 2017)*.

The Russian Federation has a complicated background regarding the HIV problem. In 2007 The Russian Federation had more people living with HIV/AIDS than
any other country in Europe, and nearly 70% of the known infections in Eastern Europe and Central Asia. Moreover, more than 80% of infected people were aged less than thirty. And despite the fact that the government took action against this issue, those measures were hindered by late recognition of the scale of the problem *(Moran & Jordaan, 2007)*. Nowadays there is a positive dynamic with regards to this problem in The Russian Federation. In 2023, the incidence of HIV decreased to 40 people per 100,000 population - this is 4.6% less than in the previous year. A decrease in the incidence of HIV infection in Russia in 2023 may be associated with the preventive effect of ART, since among the 855 thousand patients under dispensary observation in 2023, 88.3% received this therapy. *(Киселева, 2024)*.

Increasing HIV drug resistance is an important public health concern. There are several significant bodies of work that research APT treatment failures on the territory of Russia that show that the use of ART does not always lead to complete suppression of viral replication; virological efficacy in patients receiving ART in Russia was 76.7% in 2020 and 79.9% in 2021 *(Ozhmegova et al., 2024)*. Almost 40% of study patients failing ART had no DRM detected, which points to poor adherence. ART adherence is a key to the success of HIV infection suppression *(Sivay et al., 2023)*. Since the number of patients with poor ART in Russia is not insignificant, we decided to look further into this issue. Our study will provide additional insight on the significance of ART in Russia and its influence on the share of patients with undetectable viral load compared to other factors

Data usage and methodology

Initial data was taken from the site “Если быть точным”. Dataset contains information regarding the most important HIV indicators in Russia by years and cities.

Originally, data was stored as a metadata, so we had to slightly transform it, in order to make it panel data. More detailed process of transformation is described in the ETL section (Extract, Transform and Load).

After the transformation we encountered an issue, that variables, which we are interested in are expressed in numeric format. To ensure better interpretation of results and reduce the probability to stumble upon an omitted variable bias, we decided to compute our own variables, which will express:

- the number of people on ART
- people with undetectable viral load
- number of infectious disease doctors

per one HIV+ patient.

ETL process, description, EDA and IDA

ETL process

Description of initial data

Here the initial data is presented. The main problem is that all indicators are hidden in two variables - indicator_name and indicator_value.

Variable	**Description**
indicator_section	The thematic category of the indicator
indicator_name	The specific name of the indicator measured
indicator_unit	The unit in which the indicator is measured
indicator_code	Code of the indicator
object_name	The name of the geographical or administrative object
object_level	Level of the geographic object
object_oktmo	OKTMO code (identifier in Classification on Objects territory of municipal formations, used after 2014)
object_okato	OKATO code (identifier in Classification on Objects territory of municipal formations, used until 2014)
year	year
indicator_value	The numeric value of the indicator
comment	Details or the clarification about the indicator
source	The source of the data
reason_na	The reason why data is missing

Table 1. Initial data

Transformation of data

To transform the data, we decided to use pivot tables instruments in R. For the correct use of this methodics we have deleted the following variables: indicator_section, comment, source, object_oktmo, object_okato, reason_na since they were unnecessary or were duplicating the existing information. After this we united all necessary information regarding the indicators into one column.

After we have pulled out indicators with their values we faced a huge number of variable measurement issues:

1. One indicator might be measured by two slightly different methodics across the timeline and hence be represented as two different indicators. It's worth mentioning that when an indicator is measured by one method, the other method isn't used. In such situations we decided to summarize the values obtained from different methodics into one indicator.
2. For some indicators we indicated a significant number of missing values. We have deleted such indicators, since the lost values were missed by systematic issues

Then we fixed the type of values to numeric, since during the loading they were broken.

After transformation of data we decided to rename all the variables to simplify our work. You can find the renamed variables and the description in table 2.

New variables	Description	Characteristic
region	Region of the country, excluding state level	The name of the region
year	Year	Numeric variable, from 2014 to 2022
num_HIV	Number of people living with HIV	Numeric variable, includes NA
num_HIV_per100k_pop	Number of people living with HIV per 100k of general population	Numeric variable, includes NA
new_HIV	Number of people identified in a	Numeric variable, includes

	year	NA
new_HIV_per100k_pop	Number of people identified in a year per 100k of general population	Numeric variable, includes NA
num_HIV_child_from_HIV_mother	Number of children with HIV born from HIV+ mother	Numeric variable, includes NA
num_alive_child_from_HIV_mother	Number of alived born children from HIV+ mother	Numeric variable, includes NA
num_compl_pregn_women_tested_HIV	Number of pregnant women tested for HIV with completed pregnancy	Numeric variable, includes NA
num_compl_pregn_women_with_HIV_ab	Number of pregnant women with detected HIV antibodies (with completed pregnancy)	Numeric variable, includes NA
num_HIV_death	Number of people died from HIV	Numeric variable, includes NA
num_HIV_death_per100k_pop	Number of people died from HIV per 100k of general population	Numeric variable, includes NA
num_HIV_prisoners	Number of prisoners with HIV	Numeric variable, includes NA
prc_HIV_prisoners_from_all_prisoners	Percentage of HIV prisoners from all prisoners	Numeric variable, includes NA
num_inf_dis_doc	Number of infectious disease doctors	Numeric variable, includes NA
num_inf_dis_doc_per100k_pop	Infectious disease doctors per 100k of general population	Numeric variable, includes NA
num_tested_HIV_ab	Number of people from general population tested for HIV antibodies	Numeric variable, includes NA
prc_tested_HIV_ab_from_pop	Percentage of general population tested for HIV	Numeric variable, includes NA

num_registered_dispensary_obs	Number of people registered for dispensary observation (end of year, persons)	Numeric variable, includes NA
num_ART	Number of people with HIV on ART	Numeric variable
num_ART_purchased	Number of annual ART courses purchased. Minimal quantity of people, who received treatment	Numeric variable, includes NA
num_HIV_tested_viral_load	Number of people with HIV tested for viral load during the year	Numeric variable, includes NA
num_HIV_tested_CD4	Number of people with HIV tested for CD4 during the year	Numeric variable, includes NA
num_HIV_tested_ART_resist	Number of people with HIV tested for ART resistance	Numeric variable, includes NA
num_HIV_with_undetect_viral_load	Number of people with HIV with undetectable viral load in last test	Numeric variable, includes NA
prc_HIV_with_undetect_viral_load_from_ num_ART	Percentage of patients with undetectable viral load from all patients on ART	Numeric variable, includes NA
reports_supply_disruption	Reports of supply disruptions	Numeric variable, includes NA
num_removed_dispensary_obs_dtd	Number of people with HIV removed from dispensary observation due to death	Numeric variable, includes NA
num_new_HIV_per100k_tested_per100k_pop	New HIV cases per 100k tested per 100k population	Numeric variable, includes NA
num_new_HIV_men	Number of new HIV cases among men	Numeric variable, includes NA
num_new_HIV_women	Number of new HIV cases among women	Numeric variable, includes NA

num_inf_dis_doc_per_HIV_pop	Number of infectious disease doctors per all HIV+ population	Numeric variable, includes NA
prc_ART_from_HIV_pop	Percentage of people on ART from all HIV+ population	Numeric variable, includes NA
prc_HIV_with_undetect_viral_load_from_HIV_pop	Percentage of patients with undetectable viral load from all HIV+ population	Numeric variable, includes NA
cross_effect_x1x2	Cross effect of prc_ART_from_HIV_pop on num_inf_dis_doc_per_HIV_pop	Numeric variable, includes NA

Table 2. Renamed variables

As we have outlined earlier, in regression analysis we are going to consider only several, most important variables.

Our dependent variable is:

1. prc_HIV_with_undetect_viral_load_from_HIV_pop

Panel identificators:

1. year
2. region

As a variables of interest we have chosen:

1. prc_ART_from_HIV_pop
2. reports_supply_disruption
3. num_inf_dis_doc_per_HIV_pop
4. cross_effect_x1x2

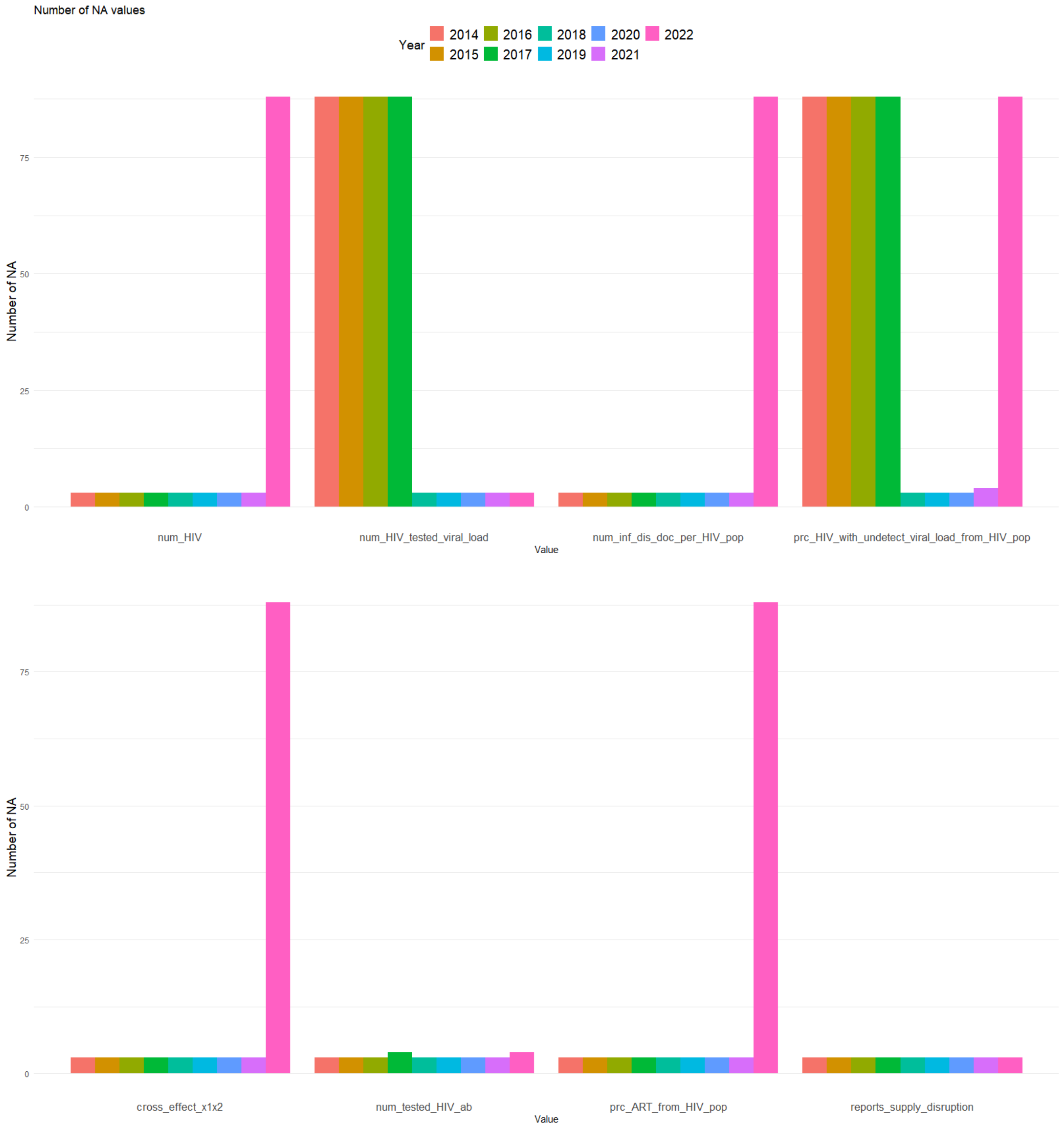
As control variables we have chosen:

1. num_HIV
2. num_HIV_tested_viral_load
3. num_tested_HIV_ab

So, all further work will be performed only with chosen regressors

\

Problem of missing values for certain years



Graph 1. Analysis on NA values

Here you can see the analysis of NA values for our chosen variables. We can see that the systematically missing values were only from 2014 - 2017 and in 2022 years. So we decided to delete these years, as the data is missing, not at random.

Explanatory data analysis

Descriptive statistic

After removing all the irrelevant variables, we computed descriptive statistics for key factors. You can find it in table 3.

The number of reported supply disruptions:

This variable reflects the number of reported disruptions in supply, which could be related to medications, tests, or healthcare delivery.
On average, there were 3.43 reported supply disruptions over the observed period (e.g., per year). The highest number of supply disruptions reported in a year was 71. This could indicate significant supply issues in certain years.

These values suggest that, during the observed period, there were both normal years and years with extreme supply disruptions.

Number of infectious disease doctors per HIV population:

This variable shows how many infectious disease doctors are available for a given HIV population, which can indicate the availability of medical services for patients.
On average, there are 0.02 infectious disease doctors per HIV patient, meaning that for every 100 people living with HIV, there are about 2 doctors.
The maximum value may indicate that, in some areas, there are 16 infectious disease doctors for every 100 people living with HIV, suggesting good medical staff availability in those areas.

The percentage of people with HIV receiving Antiretroviral Therapy, ART

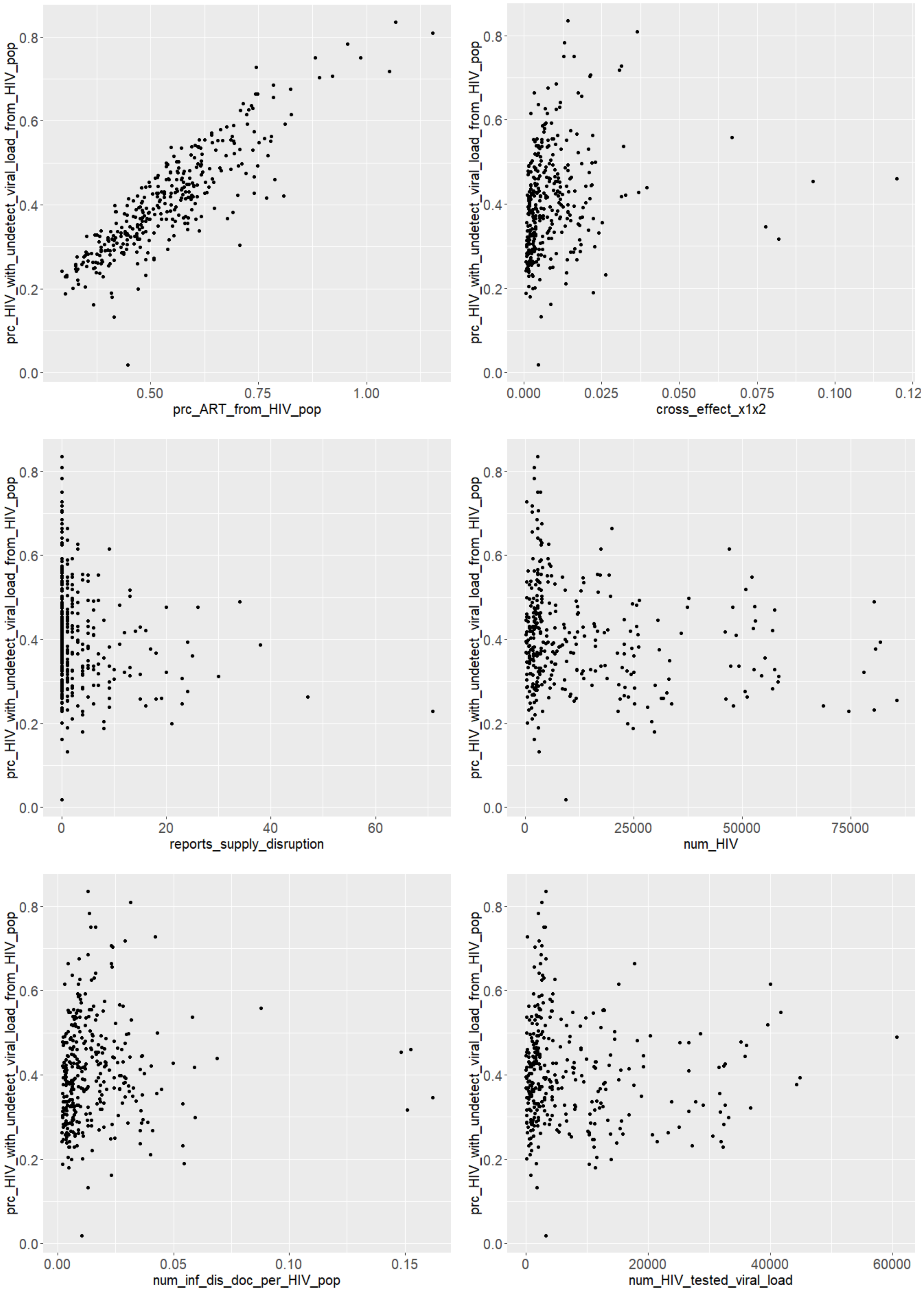
On average, 54% of people living with HIV receive ART during the observed period. This indicates that more than half of the HIV-positive population has access to ART.
In some periods or regions, only 29% of people with HIV receive ART, which could suggest lower accessibility to therapy in these areas or periods.
The maximum value exceeds 100%, which may indicate that, in some cases, the data reflects the number of people who begin therapy rather than the percentage of the total HIV population. This could also suggest data errors or special conditions of reporting.

Variable	**N**	**Mean**	**SD**	**Min**	**Max**
prc_HIV_with_undetect_viral_load_from_HIV_pop	339	0.4	0.12	0.02	0.84
year	339	2019.5	1.12	2018	2021
num_HIV	339	12715.36	17544.6	90	85570
num_inf_dis_doc_per_HIV_pop	339	0.02	0.02	0	0.16
prc_ART_from_HIV_pop	339	0.54	0.14	0.29	1.15
reports_supply_disruption	339	3.43	7.27	0	71
num_HIV_tested_viral_load	339	7382.96	9830.98	53	60734
num_tested_HIV_ab	339	467666.14	626838.19	11300	5507836
cross_effect_x1x2	339	0.01	0.01	0	0.12

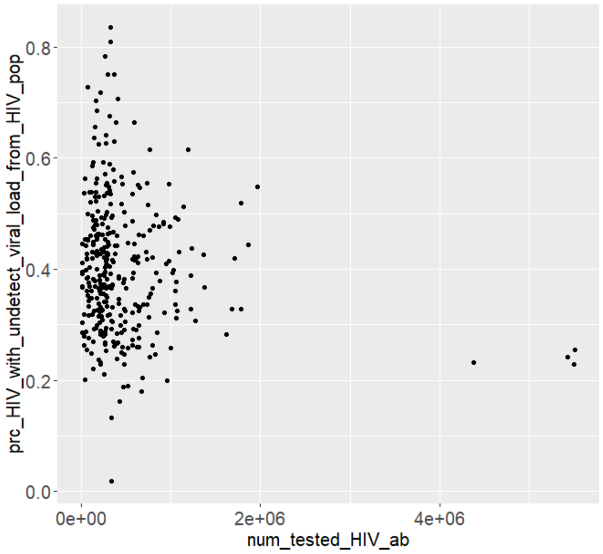
Table 3. Descriptive statistics

Visualization

Here we computed the scatter plots for the main regressors. The first scatter plot shows us that the variables have a strong positive association, for others we can't say for sure, some of them can indicate negative relation and some positive, the points are clustered at low values of X.

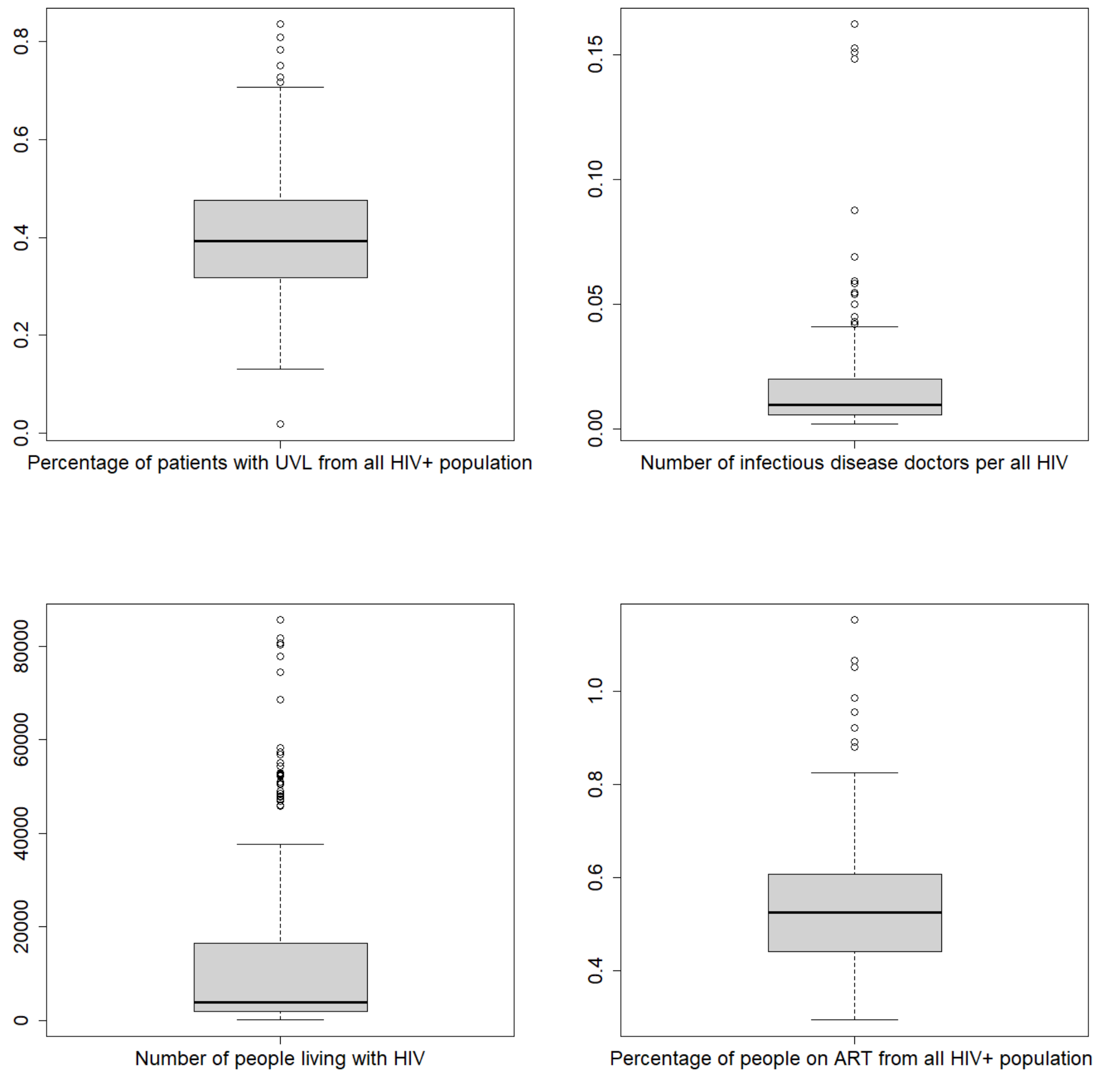


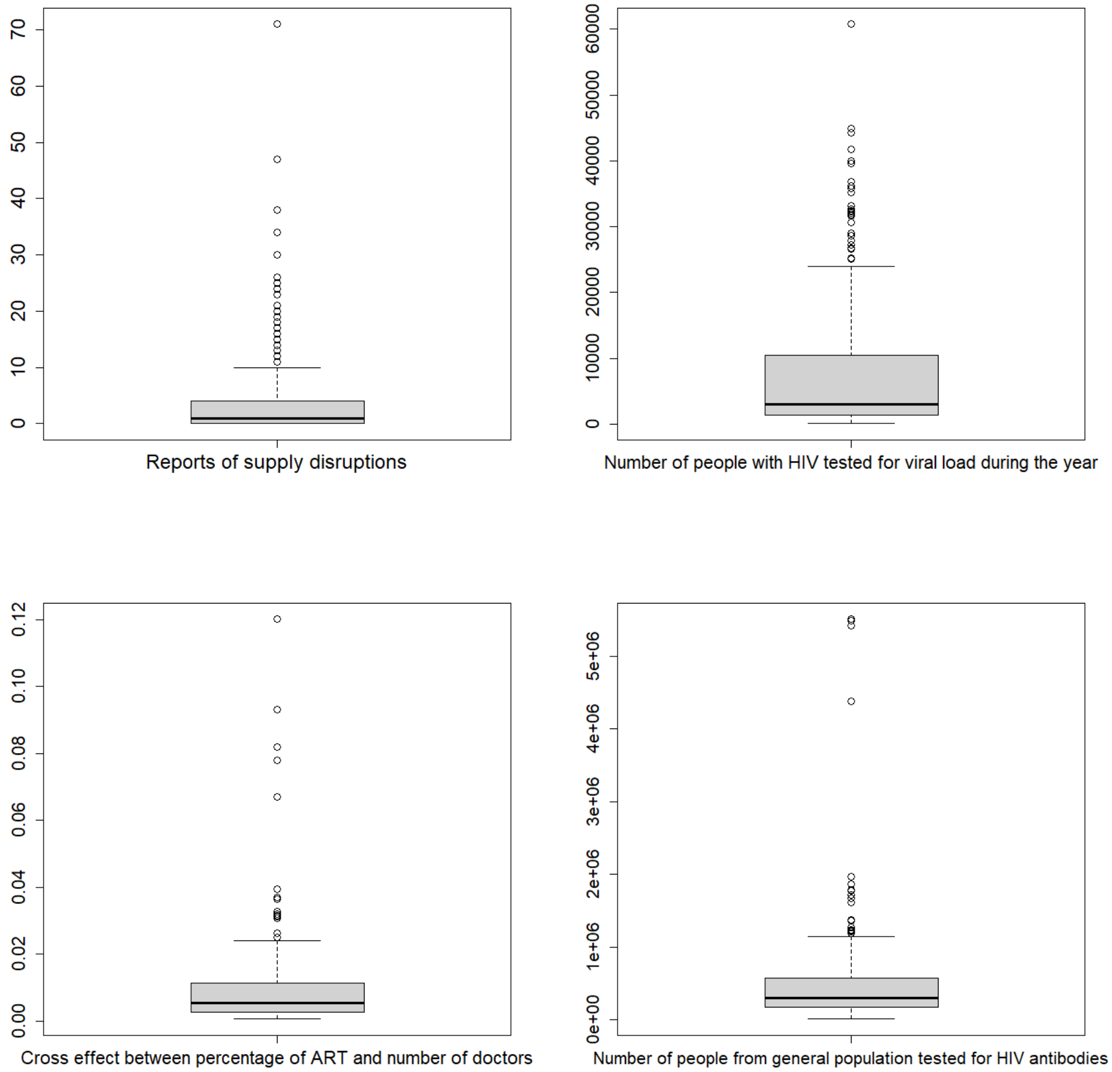
prc_HIV_with_undetect_viral_load_from_HIV_pop
0.0
0.2
0.4
0.6
0.8
0.50
0.75
1.00
prc_ART_from_HIV_pop
0.000
0.025
0.050
0.075
0.100
0.12
cross_effect_x1x2
0
20
40
60
reports_supply_disruption
0
25000
50000
75000
num_HIV
0.00
0.05
0.10
0.15
num_inf_dis_doc_per_HIV_pop
0
20000
40000
60000
num_HIV_tested_viral_load



Graph 2. Scatter plots

Here we computed the box plots for the main regressors.



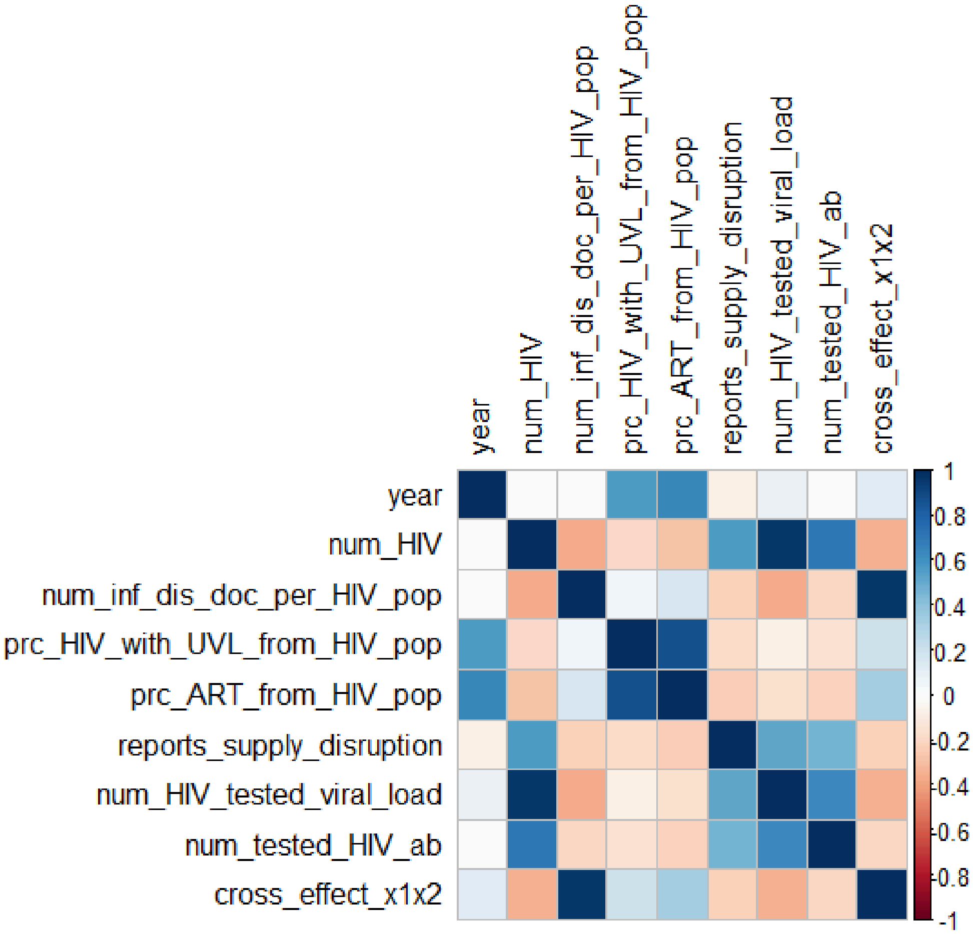


Graph 3. Box plots

We have a lot of outliers in the regressors which might lead to biased regression.

Intelligence data analysis

Correlation matrix



Graph 4. Correlation matrix

To shorten the notation UVL = undetectable viral load

num_HIV < > reports_supply_disruption
If the number of HIV people grow - the amount of medications required to treat them will also increase. Basil economic law - demand grows - supply grows, therefore we can expect that there will be more supply chain disruption

num_HIV < > num_HIV_tested_viral_load
As the number of HIV infected increases - the more tests is required to see the dynamic of the treatment

num_HIV < > num_tested_HIV_ab
Quite similar case with other correlation pairs: it is a possible consequence that the number of people from general population tested for HIV antibodies will increase with the population growth

prc_HIV_with_UVL_from_num_AR < > prc_ART_from_HIV_pop
Interesting connection as it means that the percentage of people that have weakened HIV activity enough to live normal life is increasing with the percentage of people who receive treatment. Basically, if the access to required medicine is available - it increases the number of people who can live almost without a concern about HIV

num_HIV_tested_viral_load < > num_tested_HIV_ab
If the number of tested viral load people increases it means that the need to test the dynamic of the treatment also grow (test on antibodies existence)

num_HIV_tested_viral_load < > reports_supply_disruption
Intuitively a little confusing correlation but possibly it appeared as with increasement of number people tested for viral load - the more medication required which may lead to a potential growth of supply.

Cross_effect_x1x2 < > num_inf_dis_doc_per_HIV_pop
Correlation found as number of infectionists is one of the part of the cross_effect variable

Formulation and justification of the model

Model 1 - baseline model

Since the *prc_ART_from_HIV_pop* and *num_inf_dis_doc_per_HIV_pop* are logically interrelated with each other, we decided to add a cross effect

$$\begin{aligned}&\mathit{prc\_HIV\_with\_UVL\_from\_HIV}_{it}\\ &\quad= \beta_0 + \beta_1 \mathit{prc\_ART\_from\_HIV\_pop}_{it} + \beta_2 \mathit{num\_inf\_dis\_doc\_per\_HIV\_pop}_{it} +\\ &\quad\beta_3 \mathit{cross\_effect}_{it} + \beta_4 \mathit{reports\_supply\_disruption} + u_{it}\end{aligned}$$

Model 2 - control variable *log(num_HIV)*

We decided to take the logarithm of the *num_HIV* since other dependent variables are measured in percentages. Moreover, this transformation simplifies the interpretation of coefficients

$$\begin{aligned}&\mathit{prc\_HIV\_with\_UVL\_from\_HIV}_{it}\\ &\quad= \beta_0 + \beta_1 \mathit{prc\_ART\_from\_HIV\_pop}_{it} + \beta_2 \mathit{num\_inf\_dis\_doc\_per\_HIV\_pop}_{it} +\\ &\quad\beta_3 \mathit{cross\_effect}_{it} + \beta_4 \mathit{reports\_supply\_disruption} + \beta_5 \mathit{log(num\_HIV)} + u_{it}\end{aligned}$$

Model 3 - control variable *log(num_tested_HIV_ab)*

$$\begin{aligned}&\mathit{prc\_HIV\_with\_UVL\_from\_HIV}_{it}\\ &\quad= \beta_0 + \beta_1 \mathit{prc\_ART\_from\_HIV\_pop}_{it} + \beta_2 \mathit{num\_inf\_dis\_doc\_per\_HIV\_pop}_{it} +\\ &\quad\beta_3 \mathit{cross\_effect}_{it} + \beta_4 \mathit{reports\_supply\_disruption} + \beta_5 \mathit{log(num\_tested\_HIV\_ab)} + u_{it}\end{aligned}$$

Model 4 - control variable *log(num_HIV_tested_viral_load)*

$$\begin{aligned}&\mathit{prc\_HIV\_with\_UVL\_from\_HIV}_{it}\\ &\quad= \beta_0 + \beta_1 \mathit{prc\_ART\_from\_HIV\_pop}_{it} + \beta_2 \mathit{num\_inf\_dis\_doc\_per\_HIV\_pop}_{it} +\\ &\quad\beta_3 \mathit{cross\_effect}_{it} + \beta_4 \mathit{reports\_supply\_disruption} + \beta_5 \mathit{log(num\_HIV\_tested\_viral\_load)} + u_{it}\end{aligned}$$

Model 5 - control variable *log(num_tested_HIV_ab)* and *log(num_HIV_tested_viral_load)*

$$\begin{aligned}&\mathit{prc\_HIV\_with\_UVL\_from\_HIV}_{it}\\ &\quad= \beta_0 + \beta_1 \mathit{prc\_ART\_from\_HIV\_pop}_{it} + \beta_2 \mathit{num\_inf\_dis\_doc\_per\_HIV\_pop}_{it} +\\ &\quad\beta_3 \mathit{cross\_effect}_{it} + \beta_4 \mathit{reports\_supply\_disruption} + \beta_5 \mathit{log(num\_tested\_HIV\_ab)} +\\ &\quad\beta_6 \mathit{log(num\_HIV\_tested\_viral\_load)} + u_{it}\end{aligned}$$

Model 6 - all controls together

$$\begin{aligned}&\mathit{prc\_HIV\_with\_UVL\_from\_HIV}_{it}\\ &\quad = \beta_0 + \beta_1 \mathit{prc\_ART\_from\_HIV\_pop}_{it} + \beta_2 \mathit{num\_inf\_dis\_doc\_per\_HIV\_pop}_{it} +\\ &\quad \beta_3 \mathit{cross\_effect}_{it} + \beta_4 \mathit{reports\_supply\_disruption} + \beta_5 \mathit{log(num\_tested\_HIV\_ab)} +\\ &\quad \beta_6 \mathit{log(num\_HIV\_tested\_viral\_load)} + \beta_7 \mathit{log(num\_HIV)} + u_{it}\end{aligned}$$

Model 7 - model 4 with individual fixed effects

In order to manage OV bias we have added individual fixed effects models

$$\begin{aligned}&\mathit{prc\_HIV\_with\_UVL\_from\_HIV}_{it}\\ &\quad = \beta_0 + \beta_1 \mathit{prc\_ART\_from\_HIV\_pop}_{it} + \beta_2 \mathit{num\_inf\_dis\_doc\_per\_HIV\_pop}_{it} +\\ &\quad \beta_3 \mathit{cross\_effect}_{it} + \beta_4 \mathit{reports\_supply\_disruption} + u_{it} + \beta_5 \mathit{log(num\_HIV\_tested\_viral\_load)}\\ &\quad + \lambda_i\end{aligned}$$

Model 8 - model 5 with individual fixed effects

$$\begin{aligned}&\mathit{prc\_HIV\_with\_UVL\_from\_HIV}_{it}\\ &\quad = \beta_0 + \beta_1 \mathit{prc\_ART\_from\_HIV\_pop}_{it} + \beta_2 \mathit{num\_inf\_dis\_doc\_per\_HIV\_pop}_{it} +\\ &\quad \beta_3 \mathit{cross\_effect}_{it} + \beta_4 \mathit{reports\_supply\_disruption} + u_{it} + \beta_5 \mathit{log(num\_tested\_HIV\_ab)} +\\ &\quad \beta_6 \mathit{log(num\_HIV\_tested\_viral\_load)} + \lambda_i\end{aligned}$$

Model 9 - model 6 with individual fixed effects

$$\begin{aligned}&\mathit{prc\_HIV\_with\_UVL\_from\_HIV}_{it}\\ &\quad = \beta_0 + \beta_1 \mathit{prc\_ART\_from\_HIV\_pop}_{it} + \beta_2 \mathit{num\_inf\_dis\_doc\_per\_HIV\_pop}_{it} +\\ &\quad \beta_3 \mathit{cross\_effect}_{it} + \beta_4 \mathit{reports\_supply\_disruption} + u_{it} + \beta_5 \mathit{log(num\_tested\_HIV\_ab)} +\\ &\quad \beta_6 \mathit{log(num\_HIV\_tested\_viral\_load)} + \beta_7 \mathit{log(num\_HIV)} + \lambda_i\end{aligned}$$

Regression analysis

Baseline model

```
===============================================================================
Standard-errors: Clustered (region)
OLS estimation, Dep. Var.: prc_HIV_with_undetect_viral_load_from_HIV_pop
-------------------------------------------------------------------------------
                               Model 1
-------------------------------------------------------------------------------
(Intercept)                   -0.02757 .
                              (0.01441)

prc_ART_from_HIV_pop           0.81889 ***
                              (0.02833)

num_inf_dis_doc_per_HIV_pop    0.58628
                              (0.48192)

cross_effect_x1x2             -1.99136 *
                              (0.86121)

reports_supply_disruption      0.00013
                              (0.00042)
-------------------------------------------------------------------------------
F-statistic                   70.04144 ***
p-value for F-statistic         0.00000
Num. obs.                         339
R^2 (full model)               0.76934
Adj. R^2 (full model)          0.76657
Wald-statistic                 244.81 ***
p value for Wald-statistic      6.32×10^-98
===============================================================================
*** p < 0.001; ** p < 0.01; * p < 0.05; . p < 0.1
```

With the regression we conducted the Wald test to evaluate whether a group of key predictors significantly contribute to explaining the percentage of HIV-positive individuals with an undetectable viral load. The hypotheses are almost the same for every model, with slightly different variables.

H_0: $\beta_{(\text{prc\_ART\_from\_HIV\_pop})} = \beta_{(\text{num\_inf\_dis\_doc\_per\_HIV\_pop})} = \beta_{(\text{cross\_effect\_x1x2})} = \beta_{(\text{reports\_supply\_disruption})} = 0$

H_1:At least one $\beta_{(\text{variable})} \neq 0$

As you can see in the table the p-value for Wald test is essentially zero, we reject the null hypothesis. This means the predictors prc_ART_from_HIV_pop, num_inf_dis_doc_per_HIV_pop, cross_effect_x1x2, and reports_supply_disruption are jointly significant in explaining the percentage of HIV-positive individuals with undetectable viral load.

Regression with control variables

```
==============================================================================
Standard-errors: Clustered (region)
OLS estimation, Dep. Var.: prc_HIV_with_undetect_viral_load_from_HIV_pop
------------------------------------------------------------------------------
                                  Model 2        Model 3        Model 4
------------------------------------------------------------------------------
(Intercept)                      -0.05640       -0.10441       -0.07681 .
                                 (0.04833)      (0.07912)      (0.04044)

prc_ART_from_HIV_pop              0.82474 ***    0.82124 ***    0.82282 ***
                                 (0.02828)      (0.02781)      (0.02797)

num_inf_dis_doc_per_HIV_pop       0.80307        0.70920        0.98410 .
                                 (0.59402)      (0.52070)      (0.56822)

cross_effect_x1x2                -2.17837 *     -2.06994 *     -2.32674 *
                                 (0.97205)      (0.90385)      (0.95128)

reports_supply_disruption        -0.00007       -0.00017       -0.00024
                                 (0.00052)      (0.00049)      (0.00050)

log(num_HIV)                      0.00288
                                 (0.00502)

log(num_tested_HIV_ab)                           0.00598
                                                (0.00628)

log(num_HIV_tested_viral_load)                                  0.00559
                                                               (0.00470)
------------------------------------------------------------------------------
F-statistic                      56.21959 ***   56.60589 ***   56.74142 ***
p-value for F-statistic           0.00000        0.00000        0.00000
Num. obs.                           339            339            339
R^2 (full model)                  0.76992        0.77114        0.77156
Adj. R^2 (full model)             0.76647        0.76770        0.76813
Wald-statistic                    266.2869***   261.2053***     261.5438***
p value for Wald-statistic         2.29×10^-102  2.61×10^-101     2.21×10^-10
==============================================================================
*** p < 0.001; ** p < 0.01; * p < 0.05; . p < 0.1
```

For ols_2, ols_3 and ols_4 the p-value for Wald test is very small. We reject the null hypothesis, indicating that the predictors, including log(num_HIV),log(num_tested_HIV_ab) and log(num_HIV_tested_viral_load) are jointly significant.

```
==============================================================================
Standard-errors: Clustered (region)
OLS estimation, Dep. Var.: prc_HIV_with_undetect_viral_load_from_HIV_pop
------------------------------------------------------------------------------
                                  Model 5        Model 6
------------------------------------------------------------------------------
(Intercept)                      -0.08408        0.09553
                                 (0.09152)      (0.09313)

prc_ART_from_HIV_pop              0.82269 ***    0.58691 ***
                                 (0.02813)      (0.05606)

num_inf_dis_doc_per_HIV_pop       0.94991       -0.09194
                                 (0.59961)      (0.66295)

cross_effect_x1x2                -2.29324 *     -1.12592
                                 (1.00505)      (1.06697)

reports_supply_disruption        -0.00024        0.00007
                                 (0.00050)      (0.00047)

log(num_tested_HIV_ab)            0.00112        0.00807
                                 (0.01085)      (0.01079)

log(num_HIV_tested_viral_load)    0.00479        0.17445 ***
                                 (0.00772)      (0.03330)

log(num_HIV)                                    -0.17617 ***
                                                (0.03653)
------------------------------------------------------------------------------
F-statistic                       47.28904 ***   49.56363 ***
p-value for F-statistic           0.00000        0.00000
Num. obs.                           339            339
R^2 (full model)                  0.77157        0.80508
Adj. R^2 (full model)             0.76745        0.80096
Wald-statistic                    261.0099***   53.04887***
p value for Wald-statistic         3.98×10^-101    1.63×10^-34
==============================================================================
*** p < 0.001; ** p < 0.01; * p < 0.05; . p < 0.1
```

For ols_5 and ols_6, the p-value for Wald test is almost zero. We reject the null hypothesis, confirming the significance of all predictors, including log(num_tested_HIV_ab), log(num_HIV_tested_viral_load).

Regression with individual (state) fixed effects

```
===============================================================================
Standard-errors: Clustered (region)
Fixed-effects: region: 85
OLS estimation, Dep. Var.: prc_HIV_with_undetect_viral_load_from_HIV_pop
-------------------------------------------------------------------------------
                                  Model 7        Model 8        Model 9
-------------------------------------------------------------------------------
prc_ART_from_HIV_pop               0.66058 ***    0.65936 ***    0.65938 ***
                                  (0.04240)      (0.04284)      (0.04308)

num_inf_dis_doc_per_HIV_pop        0.71800        0.73523        0.73863
                                  (0.52965)      (0.53044)      (0.58525)

cross_effect_x1x2                 -2.45488 ***   -2.48020 ***   -2.48416 ***
                                  (0.56716)      (0.56890)      (0.69566)

reports_supply_disruption         -0.00006       -0.00007       -0.00007
                                  (0.00017)      (0.00018)      (0.00018)

log(num_HIV_tested_viral_load)     0.09070 **     0.09118 **     0.09106 **
                                  (0.02689)      (0.02715)      (0.02829)

log(num_tested_HIV_ab)                            0.00645        0.00647
                                                 (0.00985)      (0.00961)

log(num_HIV)                                                     0.00057
                                                                (0.03837)
-------------------------------------------------------------------------------
F-statistic                      249.65560 ***  208.22215 ***  178.47630 ***
p-value for F-statistic            0.00000        0.00000        0.00000
Num. obs.                            339            339            339
R^2 (full model)                   0.93695        0.93700        0.93700
R^2 (within model)                 0.85723        0.85734        0.85734
Adj. R^2 (full model)              0.91441        0.91414        0.91379
Adj. R^2 (within model)            0.85437        0.85389        0.85330
Wald-statistic                    66.83683***    65.24501***    66.76066***
p value for Wald-statistic          2.41×10^-38   1.23×10^-37     3.19×10^-38
===============================================================================
*** p < 0.001; ** p < 0.01; * p < 0.05; . p < 0.1
```

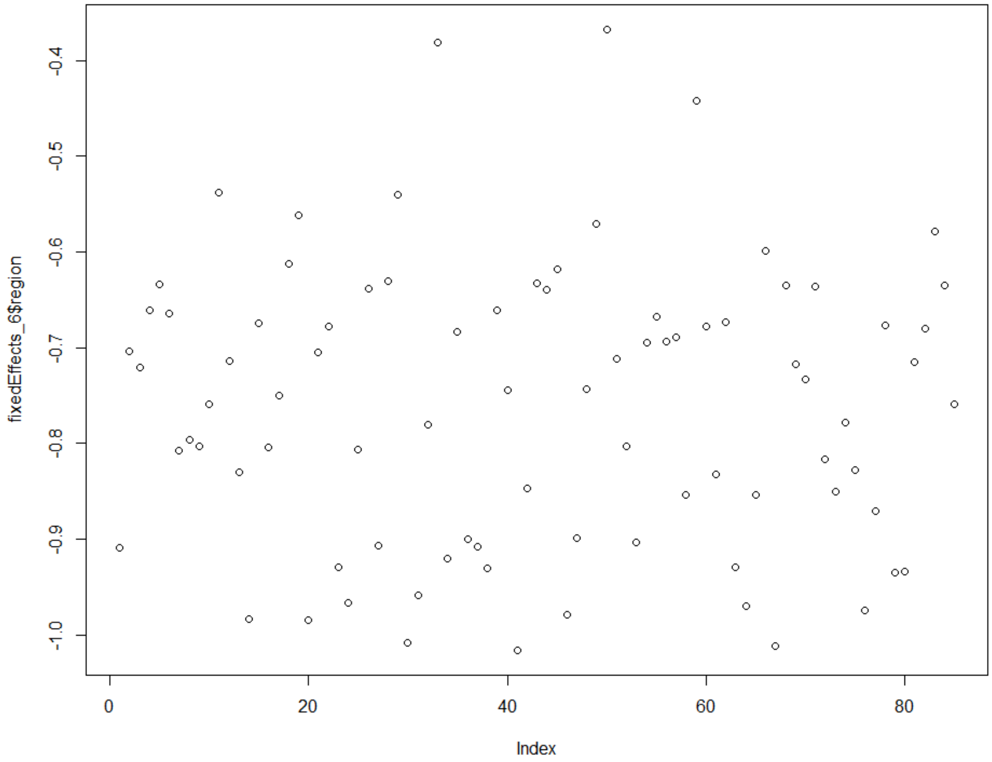
For ols_4_ife, ols_5_ife and ols_6_ife (with fixed effects), the p-value for Wald test is still significantly small. The null hypothesis is rejected, indicating that the predictors remain significant even in the most complex model with fixed effects.

Analysis of results

For the better understanding of obtained results we have conducted additional statistical and visual analysis of the obtained regression results.

Fixed effect of region

In order to clear interpretation of fixed effects of region in our most broad model 9, we have plotted a graph of fixed effects. On y-axis we have matched values of corresponding coefficients and on x-axis we matched the index of a city (basically unorder numeration to more convenient interpretation)



```
Number of fixed-effects for variable region is 85.
            Mean = -0.761     Variance = 0.0214
```

As we can see from the graph, the mean effect of the region is -0.761, which is relatively big.

Answer to research question

Let's examine our step by step

The results of model 1 suggest that the level of stability of drug supply is insignificant both econometrically and economically. The percentage of patients on ART from all patients have a significant and meaningful influence on the dependent variable, holding other things constant. The number of doctors per HIV patient is statistically insignificant, but together with the significant results of cross effect it leads to interesting conclusions.

So, the number of doctors does not influence the dependent variable alone, but it does influence it via the cross effect on percentage of patients on ART. The effect of stable drug supply is insignificant and most impact is done by the percentage of patients on ART.

Then we were aimed at testing obtained results by adding control variables. After the analysis of models 2 - model 6 we can conclude that the cross effect was not resistant to the combination of control variables. The statistical significance of the logarithm of the number of HIV patients is poorly interpreted in the real sense. It might be a result of omitted variable bias.

To tackle the OV bias we decided to check fixed individual effects in our models. The analysis of model 7 - model 9 resulted in the appearance of significant fixed region effect. This is purely logical, since every city has its own socio-economic characteristics, health care and demographic indicators. Together with this fixed effect we may conclude that the number of people tested for viral load has a significant impact on the percentage of people with UVL (dependent variable). Since the number of people tested for viral load is a control variable we may conclude that we have tackled one more omitted variable.

The general result of regression analysis is the following: percentage of people on ART is a main indicator which positively influences the percentage of “cured” people. Number of infectious disease doctors is an important variable, but it can’t influence the percentage of patients “cured” alone. Another important facts which are needed to be considered:

1. The more people are tested for UVL, the bigger the percentage of “cured” people will be.
2. The intrinsic environment in regions negatively influence the number of people who have undetectable viral load in their blood

Limitations of the study

1. People who have HIV but not included into data

Some people may not know about their disease yet - it may be the case when a person doesn't have an opportunity to test himself as the polyclinic does not have the necessary equipment. Also, we can't deny the emotional side of this issue - it may be really scary for someone to make a step forward to this disease, therefore person postpone the moment when he goes to a polyclinic.

Moreover, the incubation period lasts from 3 weeks to 3 months - in this time it is very challenging to detect HIV. As a result, the true number of HIV infected people is a little different

2. Lack of medicines

It is a widespread problem for russian realities, especially among small regions when the demand is bigger than supply. All medication is bought on the government's money and usually budget is limited, hence some people don't get treatment. Hence, we can't be sure that everybody get a proper dose of medicine, so our study may distort the real effect of ART treatment

3. Influence of migration

A certain proportion of people left the country, especially after 2022 actions, therefore some share of HIV infected left the country. We can't be sure that every HIV person is recorded in the data. Moreover, migrants may face stigma and discrimination. This can make it harder for them to access health care and support.

4. Reliability of medical records

Different standards for maintaining medical records may be used in different countries and even regions within a country. These include document formats, disease and procedure coding systems, and information storage methods (paper or electronic).

In some areas, there is no centralized database. This makes it difficult to share information between institutions. This can lead to duplication of records and confusion about the patient's data.

The human factor is also implied: errors in the completion of medical records can be the result of inadequate training of health care staff or high staff workloads. Incomplete or incorrectly completed records can make further treatment of the patient more difficult.

Medical records can be physically destroyed or lost in areas with poor infrastructure or during natural disasters (e.g. floods, fires).

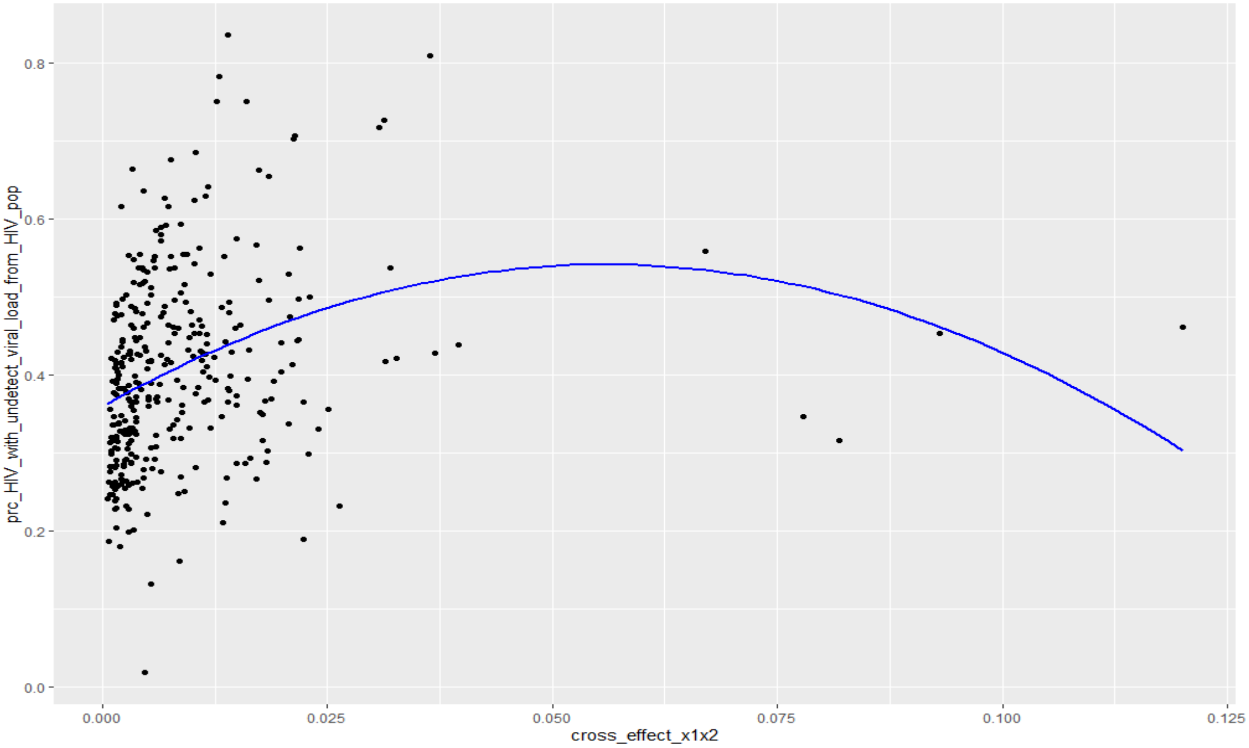
5. Multicollinearity problem and big number of outliers

In our project we also have a multicollinearity problem and it limits a little the interpretation of the coefficients, but its influence is minimized as we made all necessary actions to tackle it.

We have a lot of outliers in the regressors which might lead to biased regression. (presented on box-plots).

Possible expansion of the study

Cross effect investigation



As we can see from the graph, approximation of the single regression with a polynomial of degree 2 reveals an interesting relationship between cross_effect and dependent variable. Some additional advanced instruments might be used to understand whether the following relationship will hold in multiple regression.

While the polynomial regression shows an important link, we can still improve our study in multiple ways. One possibility is to broaden the set of explanatory variables, for example we can add socioeconomic factors and healthcare access indicators. This could provide a more comprehensive understanding of the underlying factors influencing viral suppression rate. Another potential expansion of the study is to look at how predictors change over time using more advanced time-series methods or panel data techniques, such as dynamic panel models. This could help understand the delayed effects of interventions, like ART coverage or healthcare stuffing, on viral suppression rates. Additionally, we can analyse data at a more detailed level, such as sub-regions or communities.

Conclusion

In conclusion we may say that our research has clarified the existing situation with HIV in Russia. Moreover, we outlined specific measures which can be implemented in order to improve the quality of life of HIV+ people.

However, our research has certain limitations, which require careful understanding of obtained results

Assessing team members' contributions to group work

Name	Responsibility covered
Levashov Danil	Choice of the research topic, investigation on potential datasets, construction of the regression model, application of data analysis (work in R), description of data, methods, and limitations, preparation of the defence speech
Murashko Lev	Choice of the research topic, investigation on potential datasets, preparation of the defence speech, preparation of NA values, formulation of models,
Arzamastseva Mishel	Choice of the research topic, investigation on potential datasets, descriptive statistic (work in R), description of data and different plots, possible expansion of the study, preparation of the defence speech
Kononov Analtolii	Choice of the research topic, investigation on potential datasets, finding literature regarding relevance of the research, analysis of the obtained results, participation in editing the final report, preparation of presentation, preparation of the defence speech
Doronin Nikita	Choice of the research topic, investigation on potential datasets, preparation of the defence speech, visual of the report, conceptualisation of results, organising working processes, investigating the limitations of the research

References

1. *HIV*. (2024, July 22). https://www.who.int/data/gho/data/themes/hiv-aids

2. Volberding PA, Deeks SG. Antiretroviral therapy and management of HIV infection. Lancet. 2010 Jul 3;376(9734):49-62. doi: 10.1016/S0140-6736(10)60676-9. PMID: 20609987.
3. Muhammad Haris and Rizwan Abbas. Four Decades of HIV: Global Trends, Testing Assays, Treatment, and Challenges. *Zoonoses.* 2024. Vol. 4(1). DOI: 10.15212/ZOONOSES-2023-0039
4. Zhao Y, Wu Z, McGoogan JM, Shi CX, Li A, Dou Z, Ma Y, Qin Q, Brookmeyer R, Detels R, Montaner JSG. Immediate Antiretroviral Therapy Decreases Mortality Among Patients With High CD4 Counts in China: A Nationwide, Retrospective Cohort Study. Clin Infect Dis. 2018 Feb 10;66(5):727-734. doi: 10.1093/cid/cix878. PMID: 29069362; PMCID: PMC5850406.
5. Lakshmi S, Beekmann SE, Polgreen PM, Rodriguez A, Alcaide ML. HIV primary care by the infectious disease physician in the United States - extending the continuum of care. AIDS Care. 2018 May;30(5):569-577. doi: 10.1080/09540121.2017.1385720. Epub 2017 Oct 9. PMID: 28990409; PMCID: PMC5967237.
6. Moran, D., Jordaan, J.A. HIV/AIDS in Russia: determinants of regional prevalence. Int J Health Geogr 6, 22 (2007). https://doi.org/10.1186/1476-072X-6-22
7. Киселева, А. (2024, May 1). Заболеваемость ВИЧ в 2023 году снизилась до 40 человек на 100 000 населения. Ведомости. https://www.vedomosti.ru/society/articles/2024/05/02/1034974-zabolevaemost-vich-v-rossii-snizilas?from=newsline
8. Ozhmegova, E., Lebedev, A., Antonova, A., Kuznetsova, A., Kazennova, E., Kim, K., Tumanov, A., & Bobkova, M. (2024). Prevalence of HIV drug resistance at antiretroviral treatment failure across regions of Russia. HIV Medicine, 25(7), 862–872. https://doi.org/10.1111/hiv.13642
9. Mariya V. Sivay, Lada V. Maksimenko, Tatiana M. Nalimova, Anastasiya A. Nefedova, Irina P. Osipova, Nadezda P. Kriklivaya, Mariya P. Gashnikova, Vasiliy E. Ekushov, Alexei V. Totmenin, Dmitriy V. Kapustin, Larisa L. Pozdnyakova, Sergey E. Skudarnov, Tatyana S. Ostapova, Svetlana V. Yaschenko, Olga I. Nazarova, Valery V. Shevchenko, Elena A. Ilyina, Olga A. Novikova, Aleksander P. Agafonov, Natalya M. Gashnikova. HIV drug resistance among patients experiencing antiretroviral therapy failure in Russia, 2019–2021,International Journal of Antimicrobial Agents, Volume 63, Issue 2, 2024, 107074, ISSN 0924-8579, https://doi.org/10.1016/j.ijantimicag.2023.107074.