IB Mathematics Studies Internal Assessment

How accurately can the spread of HIV/AIDS from the 1990's onwards be modelled mathematically?

Exam Session: May 2021

Table of Contents

Introduction

Human immunodeficiency virus (HIV) is a virus that attacks cells helping the body fight infections, leaving a person vulnerable to any infection or disease. If left untreated, HIV can lead to *acquired immunodeficiency syndrome* (AIDS), lowering a person's lifespan. First identified in 1981, HIV is one of the deadliest and unnoticed infections causing a persistent epidemic.[1] The early stages of the virus were brutal for the human population seeing that it was an unrecognized disease where symptoms would show until a year after the infection. These symptoms were often so sudden and lethal to infected patients. With that said, HIV infections became very hard to control among social groups until antiretroviral drugs managed to control the transmission.[2] Learning all this information made me realize how much awareness is needed about HIV. Especially within my community, where I've constantly noticed the lack of acknowledgment towards sexually transmitted infections. These diseases are a reality and we mustn't overlook them otherwise they become even more dangerous than they already are. I would like to inform myself about this to know how to take care of myself and those around me.

This investigation aims to see from which of the three selected countries I can create a more reliable model for HIV spread. By creating this model, a future prediction will contribute to preventive measures against the virus. This contribution might potentially help the targeted populations from the investigation, seeing that the spread can be controlled and the ongoing epidemic can be handled more efficiently. My prediction is that I will accurately model the spread of HIV/AIDS for the countries of Nigeria, Brazil, and Russia.

I will achieve this by using my researched data points for my model; creating a system of simultaneous linear equations in a specific order. I will conduct this process with the country of Nigeria. From the first 20 data points, I will calculate the regression line for the plotted points in a graph. This is conducted by replacing specific points into the variables of the general display of a 5th order polynomial, which is the type of equation I will be using for the creation of the model. This line is plotted along with the points to prove the accuracy of the fit between the line and the points. After conducting the same process for the remaining countries, I will find the root mean square error from each data set. The country with the smallest value will be the most accurate for a new model. The creation of a new model will

[1] HIV.gov. (2020). *What are HIV and AIDS?* [online] From: *HIV.gov.* Available at: https://www.hiv.gov/hiv-basics/overview/about-hiv-and-aids/what-are-hiv-and-aids
[2] Ibid

allow me to test two remaining data points to test the fitness of the model. The model will help me conclude on the best fit for HIV/AIDS spread overall.

Data Collection

The data from this investigation was obtained from the official University of Oxford website in the Oxford Martin School. This institution has a specific section addressing the world's data, in this case, sexually transmitted diseases such as HIV/AIDS. As mentioned before, the three countries that will be analyzed are Nigeria, Brazil, and Russia. I chose these countries because, after looking through their data, I realized how distributed all of their cases are and how different these three countries are handling the spread of HIV within their borders. I thought it would be interesting to model three countries geographically located far from each other yet dealing with the same infectious disease. Nigeria has the second-largest HIV epidemic in the world. UNAIDS estimates that two-thirds of new HIV infections in West and Central Africa occur in Nigeria.[3] HIV and AIDS epidemic in Brazil is classified as stable at a national level, however, prevalence varies geographically with higher levels in the South and Southeast of the country. Brazil's HIV epidemic is concentrated among key populations with men particularly infected. New infections have tripled among the ages 15 to 19 and doubled among the 20 and 24 ages. Brazil represents the largest number of people living with HIV in Latin America.[4] The Russian Federation has the largest HIV epidemic in Eastern Europe and Central Asia. Russia's HIV epidemic is currently concentrated among certain population groups, such as drug users and heterosexual men.[5] I chose to model between the years 1990 and 2017. I selected these years specifically with the aim of obtaining a highly significant number of data points for my model. Also to evaluate the situation at a worldwide level after approximately ten years of the disease's discovery and how this epidemic has affected infected patients chronically living with HIV/AIDS.

The following table shows the year and its corresponding number of cases for the country of Nigeria. The tables for Russia and Brazil are provided in the appendix.

[3] Avert. (2018). *HIV and AIDS in Nigeria.* [online] From: *avert.org.* Available at: https://www.avert.org/professionals/hiv-around-world/sub-saharan-africa/nigeria
[4] Avert. (2018). *HIV and AIDS in Brazil.* [online] From: *avert.rog.* Available at: https://www.avert.org/professionals/hiv-around-world/latin-america/brazil
[5] Avert. (2018). *HIV and AIDS in Russia.* [online] From: *avert.org.* Available at: https://www.avert.org/professionals/hiv-around-world/eastern-europe-central-asia/russia

Table 1. HIV Cases in Nigeria (1990 - 2017)

Year	Number of cases
1990	149092
1991	202502
1992	251950
1993	295131
1994	329387
1995	351469
1996	361579
1997	363712
1998	360255
1999	353787

Year	Number of cases
2000	347398
2001	339541
2002	327877
2003	315010
2004	303952
2005	297972
2006	296205
2007	295073
2008	294081

Year	Number of cases
2009	292538
2010	289543
2011	284575
2012	277668
2013	268894
2014	258420
2015	246271
2016	232330
2017	261614

To be able to simplify my data points and the outcomes of them, I decided to change the years into numbers from 0 to 27, representing the years 1990 to 2017. With that in mind, the table would look like this:

Table 2. HIV Cases in Nigeria (0 - 27)

Year	Number of cases
0	149092
1	202502
2	251950
3	295131
4	329387
5	351469
6	361579
7	363712
8	360255
9	353787

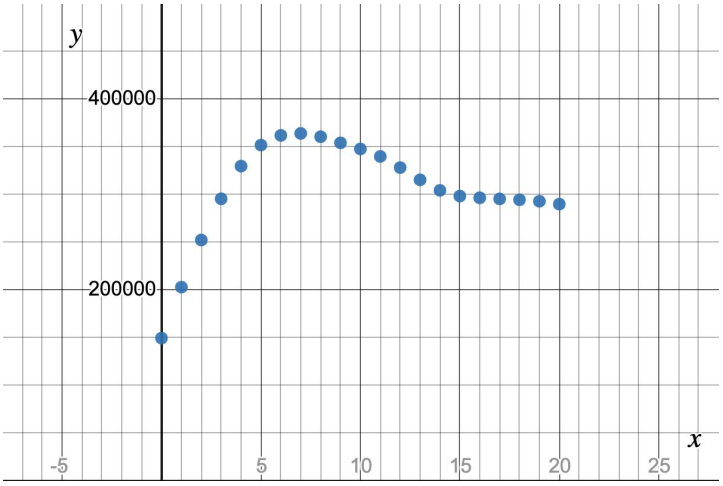
Year	Number of cases
10	347398
11	339541
12	327877
13	315010
14	303952
15	297972
16	296205
17	295073
18	294081

Year	Number of cases
19	292538
20	289543
21	284575
22	277668
23	268894
24	258420
25	246271
26	232330
27	261614

Mathematical Processes

As seen in Table 1, every country has a wide range of data regarding the number of HIV cases, which is why I've decided to mathematically model each country's data. This is done with the purpose of obtaining an accurate model to evaluate future predictions. To decide which function was best to model the data for Nigeria, I chose to plot the data on a graph.

The following graph demonstrates the points from Table 2 plotted.



Graph 1. HIV Cases in Nigeria (plotted points)

Following the visual representation of my data in Graph 1, I think a polynomial is the best type of function given that the data presented tends to constantly increase. An exponential function, for instance, wouldn't be appropriate given the fact that these functions often model behaviors that drastically increase or decrease. Because this data behaves not as drastically and it's increasingly positive, exponentials are not suitable. The equation is at a 5th degree given that a quintic function has inflection points where the graph changes shape. Because these data points create a concave up and down, inflection points are suitable for a quintic function to model this type of data.[6]

The following structure displays the composition to find a 5th degree polynomial:

$$y = ax^5 + bx^4 + cx^3 + dx^2 + ex + f$$

Consequently, six different equations are required to solve this system. These equations are obtained from the (x,y) values from Table 2. The process for obtaining one of these

[6] Glen, S. (2019). *Quintic Functions.* [online] From: *CalculusHowTo.com.* Available at: https://www.calculushowto.com/quintic-function-polynomial/

equations is to first substitute; for the year 1990 in Nigeria, the known values (0,149092) are substituted in the 5th degree polynomial structure:

$$149092 = a(0)^5 + b(0)^4 + c(0)^3 + d(0)^2 + e(0) + f$$

When simplified it looks like:

$$f = 149092$$

Following the same process for each year, the following system of equations is obtained by substituting points 0, 4, 8, 12, 16, and 20 from Table 2. I chose these points given that they are within the proper number of unknowns needed for my calculations. I also chose every four points so the year distribution can be accurately modelled through the years 1990 to 2010:

$$f = 149092$$

$$329387 = 1024a + 256b + 64c + 16d + 4e + f$$

$$360255 = 32768a + 4096b + 512c + 64d + 8e + f$$

$$327877 = 248832a + 20736b + 1728c + 144d + 12e + f$$

$$296205 = 1048576a + 65536b + 4096c + 256d + 16e + f$$

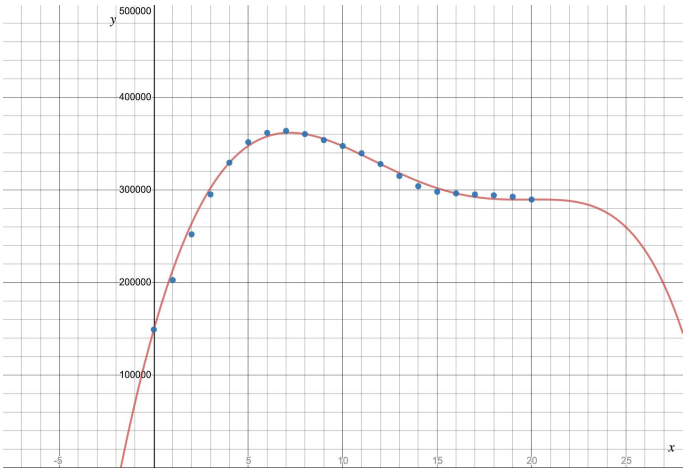
$$289543 = 3200000a + 160000b + 8000c + 400d + 20e + f$$

This system is solved through the calculator, placing all these values individually to solve the polynomial. When solved, this is the equation obtained:

$$y = -0.141756x^5 + 2.05225x^4 + 231.878x^3 - 7545.90x^2 + 71452.2x + 149092$$

where y stands for the average number of cases in an x number of years after 1990.

The following Graph 2 shows the plot for the polynomial above, attempting to model the growth in cases through the years from the data collected. As observed, a 5th degree polynomial accurately adapts to the data points.



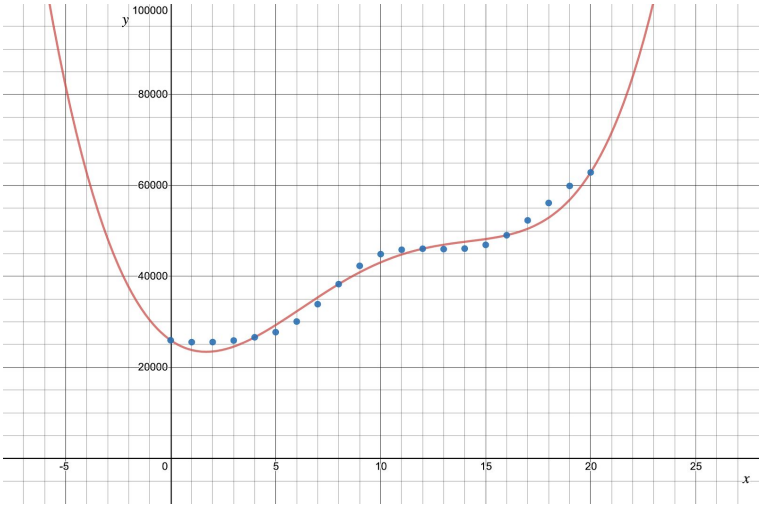
Graph 2. HIV Cases in Nigeria with Regression Line

In a similar way, I conducted the same process for Brazil and Russia.

When Brazil's system is resolved into the 5th degree polynomial, it looks like this:

$$y = 0.0206787x^5 + 1.45736x^4 - 82.1647x^3 + 1147.84x^2 - 3211.78x + 25954$$

When plotted in a graph, the equation as well as the data points, turn out like this:

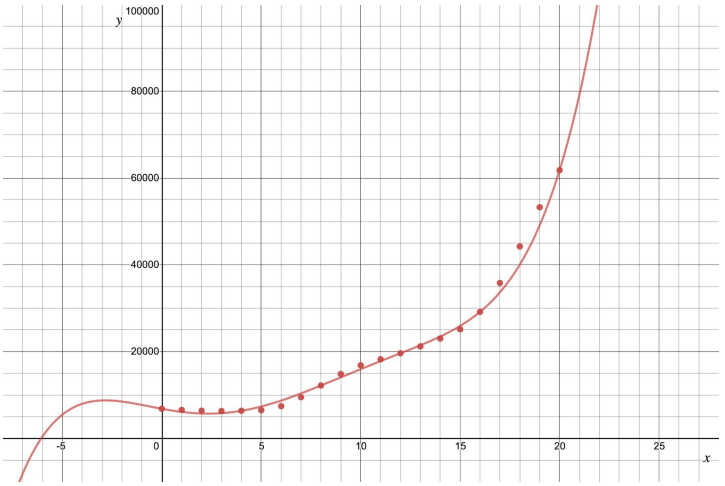


Graph 3. HIV Cases in Brazil

When Russia's system of equations is resolved, it looks like this:

$$y = 0.125513x^5 - 4.18392x^4 + 38.0501x^3 + 88.0990x^2 - 842.058x + 6815$$

And, when plotted in a graph, it looks like this:



Graph 4. HIV Cases in Russia

Visually, the graph for HIV cases in Nigeria looks more accurate but as the number of cases is a big number, I decided to measure the root mean square error (RMSE). This process finds the standard deviation of predicted errors. It measures how spread out the residuals are to determine how concentrated the data actually is. The process consisted of obtaining predicted values that were provided by calculated functions. Once I obtained these data points, I contrasted them with the actual values. If we are able to determine a certain concentration, the model will become more precise. The formula for the RMSE is:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \widehat{x}_i)^2}{N}}$$

where N stands for the number of non-missing data points, x_i stands for actual observations, and $\widehat{x}_i$ stands for estimated value from the calculated function. With this said, the process that I conducted for Nigeria was the following:

Table 3. Calculation for the RMSE for Nigeria

x_i	$\widehat{x}_i$	$x_i - \widehat{x}_i$	$(x_i - \widehat{x}_i)^2$
149092.00	149092.00	0.00	0.00
213232.09	202502.00	10730.09	115134788.49
263696.12	251950.00	11746.12	137971429.02
301927.99	295131.00	6796.99	46199100.25
329386.81	329387.00	-0.19	0.04
347529.92	351469.00	-3939.08	15516359.12
357795.87	361579.00	-3783.13	14312080.16
361587.41	363712.00	-2124.59	4513869.92

360254.49	360255.00	-0.51	0.26
355077.22	353787.00	1290.22	1664677.97
347248.90	347398.00	-149.10	22230.81
337858.97	339541.00	-1682.03	2829241.74
327876.01	327877.00	-0.99	0.98
318130.77	315010.00	3120.77	9739192.91
309299.09	303952.00	5347.09	28591360.77
301884.94	297972.00	3912.94	15311130.75
296203.41	296205.00	-1.59	2.54
292363.64	295073.00	-2709.36	7340647.87
290251.89	294081.00	-3829.11	14662075.73
289514.48	292538.00	-3023.52	9141649.00
289540.80	289543.00	-2.20	4.84

This process determined how close predicted values are from actual values so that means that the lower the value, the better the closeness. The result of Nigeria's process for the RMSE is 4488. This was obtained by rooting the result of the squared value in the last column of Table 3. The same process is repeated for Brazil and Russia and given in Table 4 below.

Table 4. Calculation for the countries' RMSE

Country	**RMSE**
Brazil	1588
Russia	1480
Nigeria	4488

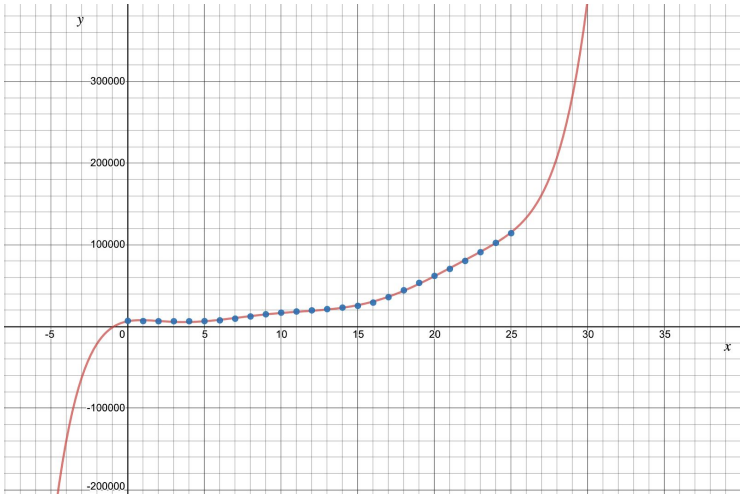
For the case of Russia, compared to the rest of the countries, it presented the lowest RMSE, as seen in Table 4, meaning that it's the most accurate model of them all. In light of this information, the creation of a mathematical model from Russia's data points felt appropriate in order to obtain accurate predictions. To be able to do this, I plotted the remaining data points I had already collected and left aside the last two so I can use them as a test for the new model.

I originally used 20 data points to create my first model because I wanted to see with which country I could then use more years to improve the model as I went through. Therefore, as

Russia had the smallest RMSE, I decided to add five more points to my data to see if I can create a more accurate model for this data set. I decided to use a polynomial function at a 7th degree rather than a 5th degree, just like I've been working with. A 7th-degree polynomial was a better fit given that a higher degree gives more accuracy. As I wasn't able to do this on my calculator due to too many unknowns, I decided to use Google sheets instead. I used sheets for the scatter graphs and to create the best regression formula for the data points. This is the formula they had with the values rounded to eight significant figures:

$$y = 0.0018580989x^7 - 0.16593321x^6 + 5.7514819x^5 - 96.757112x^4 + 810.96964x^3 - 2964.2406x^2 + 3535.5300x + 6307.0314$$

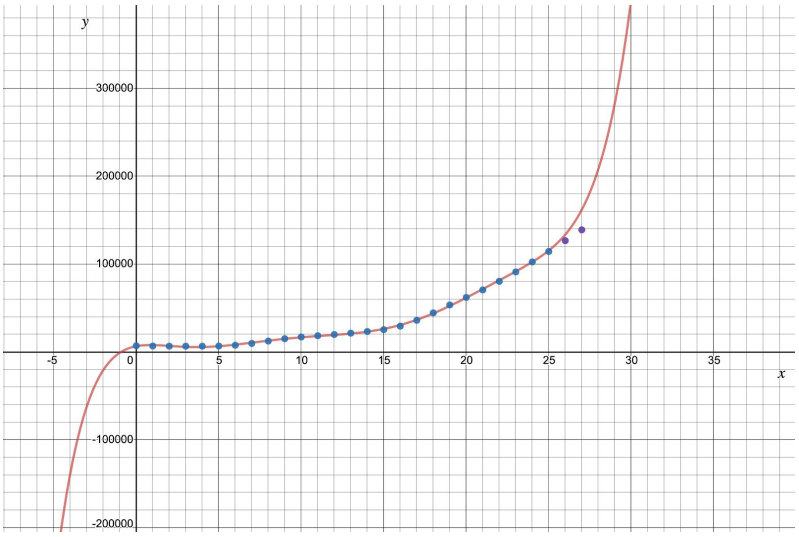
When this equation is plotted along with its data points, the graph shapes like this:



Graph 5. New Model for Russia

Seeing that this is a new model, calculating its RMSE will help me determine the level of accuracy. I conducted the same process that I previously performed for the three countries and the results for Russia's new model RMSE is 712 which shows a decrease in the value from the previous Russia RMSE's calculation, which was 1480. This decrease means that the error margin has been reduced from the new model, making it more accurate. Furthermore, I decided to test for the two values that were left, which are the years of 2016 and 2017.

When these two points are tested, the model looks like the following:



Graph 6. Remaining Data Points Plotted into New Model

where the two purple dots represent the years 2016 and 2017 of HIV cases in Russia. This will be discussed in the interpretation in the next section. As my original question was to see how accurately I could model my data, I chose to work out the percentage error of 2016 and 2017. I calculated this using the following formula:

$$\delta = \left| \frac{v_A - v_E}{v_E} \right| \cdot 100\%$$

where δ refers to the percent error, v_A represents the actual value observed, and v_E represents the expected value.

By successfully substituting the values into the formula, my calculations were: 5.26% for 2016 and 14.2% for 2017. The importance of this will be discussed further in the interpretation section.

Interpretation of Results and Conclusions

When considering Table 4, each country presents a different value for its RMSE. This difference is given because of the increase or decrease in HIV cases through the specifically chosen timeframe. A significant increase, for instance, causes the data points to become more separated from each other and to have an impact on the RMSE. Since these values determine approximately how many patients will become infected in the future, the smaller they are, the more significant the model is. This is because the values are calculated between predicted and actual, so smaller outcomes indicate there is a higher relevance. When I plotted the equations to be able to obtain predicted values, I noticed they are not as significantly different as the actual values. Meaning that the predictions elaborated by that

model are highly accurate. We can see in graphs such as Graph 5, how all data points fit the line of regression. This indicates the calculations for the model are functional and ready to be tested.

For the new model that uses a 7th order polynomial, the RMSE is 712. When this value is compared, it's less than other countries which means that I've improved my model by reducing the error. However, this model comes with its own limitations. I had initially reserved the data for 2016 and 2017 to test this data. I found that using this new 7th order model, the percentage errors were 5.26% and 14.2% respectively, between the real values and the values calculated by the model. A reasonably low error margin from both of these data points demonstrates that predicted values from the model are not too far from actual values. However, one reason for there being a percentage error between these values is because the values are extrapolated as they are not within the range of data used to create the model. There is still a slight difference between the line and where the two points fit. This observation might lead to the fact that the model is not as precise to determine the future spread of HIV/AIDS in the three countries.

Because the quantities for my new model's formula were too large, I decided to plot the equation with eight significant figures. I realized that if I used the equation at three significant figures, the accuracy for the RMSE calculation wouldn't be as precise as required. Therefore, I increased the significance at an appropriate level to reduce the error. What I noticed with this setback was that rounding is a very important aspect to consider when creating mathematical models. A slight error would completely change the expected predictions.

As seen in the last calculations of the Mathematical Processes section, there is a reduction in the RMSE for the new model. This indicates that the error of the model's prediction is reduced which makes it more efficient. As a result, I can conclude that I created a realistic model that portrays the HIV/AIDS epidemic. This means the model created has a low percentage error of failing at making predictions. Despite there being a percentage present, the model still holds a purpose. Said purpose is important to prevent a future spread of the virus.

References

1. Avert. (2018). *HIV and AIDS in Brazil.* [online] From: *avert.rog.* Available at: https://www.avert.org/professionals/hiv-around-world/latin-america/brazil
2. Avert. (2018). *HIV and AIDS in Nigeria.* [online] From: *avert.org.* Available at: https://www.avert.org/professionals/hiv-around-world/sub-saharan-africa/nigeria
3. Avert. (2018). *HIV and AIDS in Russia.* [online] From: *avert.org.* Available at: https://www.avert.org/professionals/hiv-around-world/eastern-europe-central-asia/russia
4. Glen, S. (2013). *Mean Squared Error: Definition and Example.* [online] From: *StatisticsHowTo.com.* Available at: https://www.statisticshowto.com/mean-squared-error/#:~:text=General%20steps%20to%20calculate%20the,original%20to%20get%20the%20erro
5. Glen, S. (2019). *Quintic Functions.* [online] From: *CalculusHowTo.com.* Available at: https://www.calculushowto.com/quintic-function-polynomial/
6. Glen, S. (2020). *RMSE: Root Mean Square Error.* [online] From: *StatisticsHowTo.com*. Available at: https://www.statisticshowto.com/probability-and-statistics/regression-analysis/rmse-root-mean-square-error/
7. HIV.gov. (2020). *What are HIV and AIDS?* [online] From: *HIV.gov.* Available at: https://www.hiv.gov/hiv-basics/overview/about-hiv-and-aids/what-are-hiv-and-aids
8. Roser, M & Ritchie, H. (2019). *HIV/AIDS.* [online] From: *ourworldindata.org.* Available at: https://ourworldindata.org/hiv-aids
9. Saudi Journal of Biological Sciences. (2018). *HIV epidemiology in Nigeria.* [online] From: *ncbi.nlm.nih.gov.* Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5937013/

Appendix

Table 5. HIV Cases in Brazil (1990 - 2017)

Year	Number of cases	Year	Number of cases	Year	Number of cases
1990	25954	2000	44900	2009	59905
1991	25538	2001	45866	2010	62885
1992	25555	2002	46086	2011	65381
1993	25899	2003	46007	2012	68103
1994	26608	2004	46118	2013	71008
1995	27734	2005	46941	2014	74042
1996	30075	2006	49058	2015	77166
1997	33894	2007	52326	2016	80289
1998	38300	2008	56156	2017	83333
1999	42324				

Table 6. HIV Cases in Russia (1990 - 2017)

Year	Number of cases	Year	Number of cases	Year	Number of cases
1990	6815	2000	16835	2009	53299
1991	6543	2001	18238	2010	61828
1992	6368	2002	19621	2011	70399
1993	6304	2003	21195	2012	80172
1994	6349	2004	23017	2013	90891
1995	6483	2005	25159	2014	102270
1996	7414	2006	29161	2015	114186
1997	9488	2007	35830	2016	126473
1998	12174	2008	44254	2017	138843
1999	14840				