

UDC 519.2

DOI: <https://doi.org/10.17721/1812-5409.2026/1.4>

Olga VASYLYK, DSc (Phys. & Math.), Assoc. Prof.
ORCID ID: 0000-0002-0880-3751
e-mail: vasylyk@matan.kpi.ua

National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukraine

Yehor OKUNIEV, Master's Student
ORCID ID: 0009-0009-3695-490X
e-mail: egorky96@gmail.com

National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute",
Kyiv School of Economics, Kyiv, Ukraine

MODELLING HIV/AIDS EPIDEMIC IN UKRAINE AS A CONTINUOUS-TIME MARKOV PROCESS

Ukraine remains among the countries with the highest rates of HIV/AIDS spread globally. In this study, we fit a continuous-time Markov model to describe the progression of HIV to AIDS and ultimately to death for an individual and so obtain the means to model the epidemic on the global level. Our analysis reconstructs the information lost due to left- and interval-censorship of empirical data. Our methodology integrates survival analysis, maximum likelihood estimation, Monte Carlo simulation, and the construction of confidence intervals. The results provide insights into the dynamics of the epidemic in Ukraine and can further be used in actuarial research of HIV/AIDS as an insurance risk.

Keywords: survival models, multiple decrements, Markov models, simulation, stochastic process.

AMS 2020 classification: 60G15, 60G22, 60G51, 68U20.

Introduction

The HIV/AIDS epidemic remains one of the most pressing challenges facing Ukraine's healthcare system (Kruglov et al., 2008; Barnett et al., 2000). According to international reports, Ukraine is among the countries with the highest rates of HIV/AIDS transmission (Miranda et al., 2021; Parczewski et al., 2025). To ensure effective resource allocation, strategic planning, and informed healthcare policy-making, a formalized approach to modeling the epidemic at the individual level is essential.

The use of discrete multi-state models to describe the progression of chronic diseases has long been a standard practice, with Markov models being particularly prominent (Johnson, Kemp, & Kotz, 2005). In this study, we first fit a discrete-time Markov chain to represent the progression of HIV/AIDS at the individual level and then extend this framework to a continuous-time Markov process.

A major challenge in the modeling process is the presence of left- and interval-censoring in the empirical data (Cox, 1972). After fitting the continuous-time Markov process, we address how the information lost due to censoring is reconstructed by the model's design.

The objective of this paper is to develop a statistical model of the HIV/AIDS epidemic that captures both local and broader dynamics, taking into account the epidemiological context of Ukraine. Furthermore, we conduct simulation using *msm* package in R (Jackson, 2007) and explore the application of the derived model in the actuarial analysis of HIV/AIDS as an insurance-related risk.

1. Data sources introduction and preparation

This study uses aggregated region-level data from the State Enterprise "Center of Public Health" of the Ministry of Health of Ukraine (HIV/AIDS.Data, 2025), covering the period from January 2014 up until the present time. The dataset includes the following information for each region and month:

- number of newly reported cases of HIV,
- number of newly reported cases of AIDS,
- number of AIDS-related deaths,
- total number of people with HIV/AIDS under medical supervision at month-end,
- corresponding rates per 100 000 population.

Due to the nature of data collection, these data are left- and interval-censored. The presence of statistics calculated per 100 000 people and the lack of explicit population data imply that these values were estimated. Additionally, we used data on general mortality rates at the region level, sourced from the website of the State Statistics Service of Ukraine, covering the period from January 2015 to January 2022 (State Statistics Service of Ukraine, 2022). Information for later dates is unavailable due to Russia's full-scale invasion of Ukraine.

The empirical data from the sources were stored in a format unsuitable for direct import into Microsoft Excel, so Optical Character Recognition (OCR) technology was employed. For this task, ABBYY software was used, as it specializes in OCR of documents resembling Excel spreadsheets. However, this approach had several limitations. Notably, the software occasionally misinterpreted the digits 0 and 1 as the letters "o" and "i", respectively; introduced rare and sporadic data corruption; and interpreted center-aligned symbols as entries containing multiple spaces.

In order to restore data lost due to shortcomings of OCR, we manually removed extra spaces and applied global character substitution of the form $o \rightarrow 0$.

Afterward, where possible, we derived the number of people who died due to causes different from AIDS (CDA). Later, we made a reasonable assumption that the number of people who die due to CDA within a certain group (e.g., people with HIV) is proportional to the number of people who died due to CDA in the general population (P).

This allowed us to estimate the number of people who died due to CDA in the following groups: people who are undiagnosed (U), people living with HIV (H), and people living with AIDS (A). It should be emphasized that undiagnosed includes individuals with HIV or AIDS who are unaware of their condition, which is realistic, given that HIV may be undetectable for up to half a year after contraction and asymptomatic for years thereafter.

Further, we will use the following right subscripts:

- $_$ - denoting the total number of people in a certain group,
- N - denoting the increment in the size of a certain group,
- D - denoting the number of people who died in a certain group,
- A - denoting death due to AIDS,
- CDA - denoting death due to CDA,
- $.100k$ - calculated per 100 000 people.

Additionally, we use the left subscript “ t ” to specify the moment in time. For example, the total number of people who died due to CDA at time point 4 can be written as ${}_4P_{D,CDA}$. Thus, our estimate for the number of people who died due to CDA within an arbitrary group G can be expressed as:

$${}_tG_{D,CDA} = {}_tP_{D,CDA} \cdot \frac{{}_tG}{{}_tP}.$$

As previously mentioned, statistics per 100 000 people were estimated and thus implicitly carry information about estimated population sizes. This allows us to recover estimates for group sizes using the formula:

$${}_tG_{_,100k} = \frac{{}_tG_{_,100k}}{{}_tP} \cdot 100\,000,$$

where $_$ stands for any statistic where the calculation of the corresponding statistic per 100 000 people is present.

Lastly, to address sporadic data corruption, we applied the following formulae to reconstruct the data:

$$\begin{aligned} {}_tH &= {}_{t-1}H + {}_tH_N - {}_tH_{D,CDA}; \\ {}_tA &= {}_{t-1}A + {}_tA_N - {}_tA_{D,CDA} - {}_tA_{D,A}. \end{aligned}$$

It should be noted that the number of progressions and regressions between HIV and AIDS is not accounted for, since currently, all people with HIV/AIDS under medical supervision receive antiretroviral therapy (ART), which eventually reverts AIDS back to HIV and puts HIV into remission. This makes such transitions unlikely and impossible to estimate with the given empirical data.

2. Designing the model

Natural choice of states would be Undiagnosed (U), has HIV (H), has AIDS (A), and Dead (D). Intuitively, one might want to connect these states in the way the virus progresses naturally, i.e., $U \rightarrow H \rightarrow A$, with all of them connected to D . However, that model is not realistic, due to several reasons:

- The widespread use of ART treatment allows transition $A \rightarrow H$,
- People under medical supervision almost never transition from H to A (Raboud et al., 2002),
- Most people in state A are “late presenters,” i.e., individuals who are diagnosed in the late stages of the infection (Miranda et al., 2021; Parczewski et al., 2025).

So, a more realistic picture (Fig. 1) of the Markov model would be:

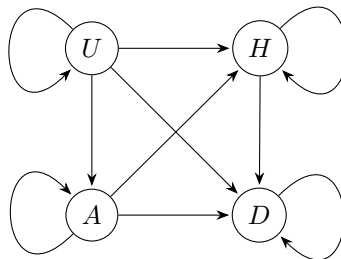


Fig. 1. Detailed model (inestimable)

The only problem with that model is that it is not possible to estimate the transition probability for $A \rightarrow H$ from the empirical data. Therefore, for the sake of further research, we introduce a new state H/A , for which all the transition probabilities can be estimated from empirical data. In this model, Markov chain (Fig. 2) and the monthly transition probability matrix take the following forms:

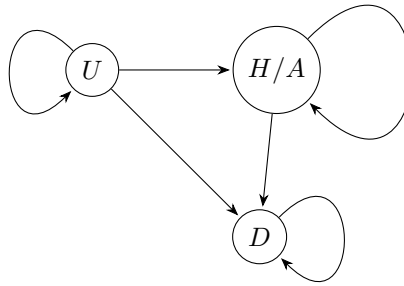


Fig. 2. Simplified model (estimable)

$$\begin{bmatrix} 1 - p_{U,H/A} - p_{U,D} & p_{U,H/A} & p_{U,D} \\ 0 & 1 - p_{H/A,D} & p_{H/A,D} \\ 0 & 0 & 1 \end{bmatrix}.$$

It should be noted, that reversions $A \rightarrow D$ are not represented in ${}_tH_N$.

Lastly, even though state U includes people with HIV/AIDS, it does not contribute to the number of deaths caused by AIDS, since most AIDS-related deaths occur among individuals who are diagnosed with AIDS at a critical stage and die before their immune system can be restored through ART.

With a discrete time-homogeneous Markov model in mind, we consider that the time until a certain transition is geometrically distributed, and the total number of transitions during a month is binomially distributed (with the same p , and different n for each month). It is intuitive to derive ML-estimator for p :

$$\begin{aligned} L(p) &= \prod_{i=1}^n C_{n_i}^{x_i} p^{x_i} (1-p)^{n_i-x_i}, \\ \mathcal{L}(p) &= \ln L(p) = \sum_{i=1}^n \ln C_{n_i}^{x_i} + x_i \ln p + (n_i - x_i) \ln(1-p), \\ \mathcal{L}'(p) &= \frac{1}{p} \sum_{i=1}^n x_i - \frac{1}{1-p} \sum_{i=1}^n (n_i - x_i) = \frac{1}{p(1-p)} \left(\sum_{i=1}^n x_i - p n_i \right) = 0, \\ \therefore p &= \frac{\sum x_i}{\sum n_i}. \end{aligned}$$

Thus, estimators for transition probabilities of interest are:

$$\begin{aligned} p_{U,H/A} &= \frac{\sum {}_tH_N + \sum {}_tA_N}{\sum {}_tU}, \\ p_{U,D} &= \frac{\sum {}_tU_{D.CDA}}{\sum {}_tU}, \\ p_{H/A,D} &= \frac{\sum {}_tH_{D.CDA}}{\sum {}_tH} + \frac{\sum {}_tA_{D.CDA} + \sum {}_tA_{D.A}}{\sum {}_tA}. \end{aligned}$$

With the fitted model, we can address the problem of censorship. Firstly, it is easy to find a continuous-time Markov chain that, when discretized by months, turns into the discrete time-homogeneous Markov model we have just fitted to the empirical data. In order to obtain the transition matrix for the discrete time-homogeneous Markov chain from a continuous-time Markov chain with generator Q , we use the formula $P = e^{tQ}$; thus, to reconstruct the generator function we use $tQ = \ln P$.

With the interval censorship removed from the distribution, the best way to address left-censorship is to shift all the information by a period of 6 months to account for the undetectability of HIV in the early stages. With exponential sojourn times, accounting for the starting age of observation does not yield any improvements, due to the lack of memory property of the exponential distribution.

With the individual epidemic Markov model derived, it is straightforward and described in (Johnson, Kemp, & Kotz, 2005) how to obtain the population epidemic Markov model, given the individual model generator Q and the initial population by groups, the mean vector and covariance matrix are expressed as follows:

$$x(t) = x(0)e^{tQ}, \quad (1)$$

$$\mathcal{C}(t) = \text{Cov}[x(t)] = \text{Diag}(x(0)e^{tQ}) - e^{tQ^\top} \text{Diag}(x(0))e^{tQ}. \quad (2)$$

3. Simulation testing

Using results from the previous section, we obtain the following estimate for the discrete-time transition matrix P :

$$\begin{bmatrix} 0.998931 & 3.361732e-05 & 0.001035343 \\ 0 & 9.450082e-01 & 0.054991825 \\ 0 & 0 & 1 \end{bmatrix}.$$

Estimate for the continuous counterpart of the same individual-level process generator Q is:

$$\begin{bmatrix} -0.001069532 & 3.459568e-05 & 0.001034936 \\ 0 & -0.056561701 & 0.056561701 \\ 0 & 0 & 0 \end{bmatrix}.$$

The estimated initial population vector is:

$$\hat{x}(0) = (43192218, 171418, _).$$

The number of dead individuals is set to zero, since we are interested in modeling only the number of people living with HIV or AIDS.

Expected vector after 1 month of observation:

$$\hat{x}(1) = (43145390, 153817, _).$$

Absolute percentage differences from observed data at $t = 1$:

$$\delta x(1) = (0.02018511, -0.1086147, _).$$

Expected vector after 3 months of observation:

$$\hat{x}(3) = (43051887, 124446, _).$$

Absolute percentage differences from observed data at $t = 3$:

$$\delta x(3) = (-0.03802216, -0.2782137, _).$$

It can be seen that forecasts based on the initial time of observation tend to underestimate the total number of individuals living with HIV/AIDS. This is caused by two factors: the assumption of exponential sojourn times, and the inability to account for the starting age at the beginning of the observation period. Estimate for the population process covariance matrix $\mathcal{C}(3)$:

$$\begin{bmatrix} 1795865.4 & -119254.9170 & -179489.9249 \\ -119254.9 & 124101.3380 & -518.8347 \\ -179489.9 & -518.8347 & 186522.1861 \end{bmatrix}.$$

We used the *msm* library for R (Jackson, 2007) to simulate the population process. To achieve this, we conducted multiple individual-level simulations and aggregated the resulting data into a simulated population-level process. This allowed us to obtain confidence intervals for the population-level values at each moment in time. By performing multiple simulations, we derived empirical confidence intervals for the population process at each time point. The results are shown on the Fig. 3-5

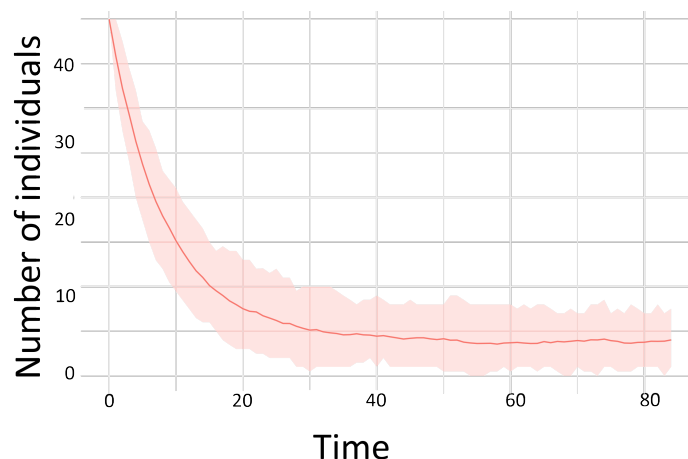


Fig. 3. CI for the number of people with HIV/AIDS with population 10000 and proportional counts

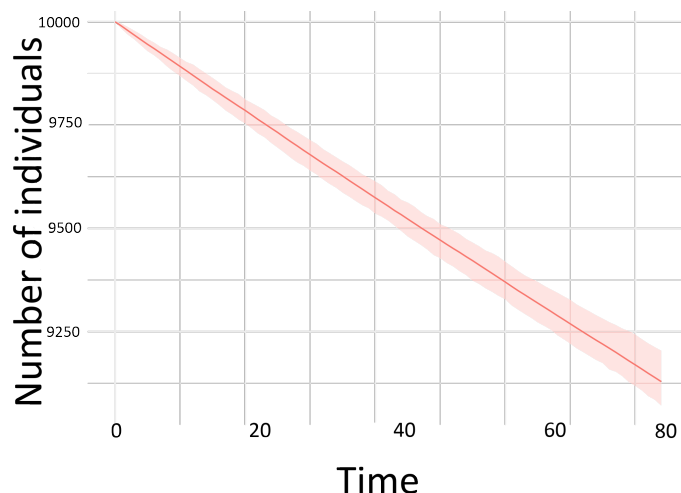


Fig. 4. CI for the number of undiagnosed people with or without HIV/AIDS with population 10000 and proportional counts

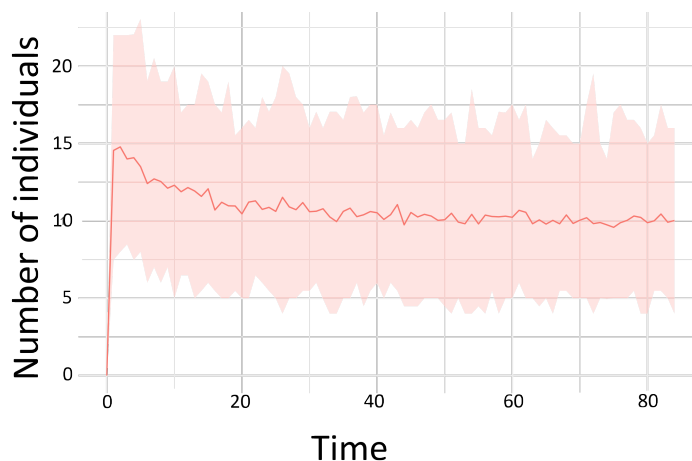


Fig. 5. CI for the number of dead people with or without HIV/AIDS with population 10000 and proportional counts

Within this model, we did not account for population growth, since most people who tested positive for HIV or AIDS were aged 15 or older. The number of parenteral sources of infection, as can be seen in (HIV/AIDS.Data, 2025), is negligible. Additionally, as stated in (Fowler, Flynn, & Aizire, 2018), children born to women living with HIV/AIDS undergo prophylaxis, and less than 1 % of such children contract HIV; thus, accounting for newborns or other parenterally infected individuals (Kleinman, & Busch, 2000) is unnecessary during the observation period.

Discussions and conclusions

This study explored the dynamics of the HIV/AIDS epidemic in Ukraine using a continuous-time Markov process model fitted to the empirical data. The model assumes that the time to transition between states U, H/A, and D is exponentially distributed, which carries an inherent assumption about memoryless progression. While analytically tractable, this assumption introduces certain limitations in capturing temporal variability in transition behavior.

The forecasts generated by the model tend to underestimate the future number of individuals living with HIV/AIDS. This bias is attributable to the non-constant nature of transition intensities over time, as shown in the Fig. 6:

In particular, the observed variation suggests that the assumption of homogeneous transition rates is too restrictive for capturing the evolving epidemiological landscape-driven by factors such as changes in testing practices, treatment accessibility, and behavioral patterns.

A natural methodological improvement would involve changing the model to model the individual level as a semi-Markov process, in which sojourn times follow more flexible distributions, such as the Weibull distribution. This extension would relax the memoryless assumption and allow for modeling duration-dependent transitions. Importantly, it would also enable more accurate inference about long-term epidemic trajectories.

The semi-Markov approach has a discrete-time counterpart in time-inhomogeneous Markov chains, which could be used to estimate a new model's parameters using LSE estimation on available data. Moreover, such models offer a framework for incorporating the starting age of individuals, thereby addressing the left-censoring issue present in the current analysis. Integrating age-structured entry would allow the model to more faithfully represent the natural history of HIV progression in a population with varied entry times into the observation window.

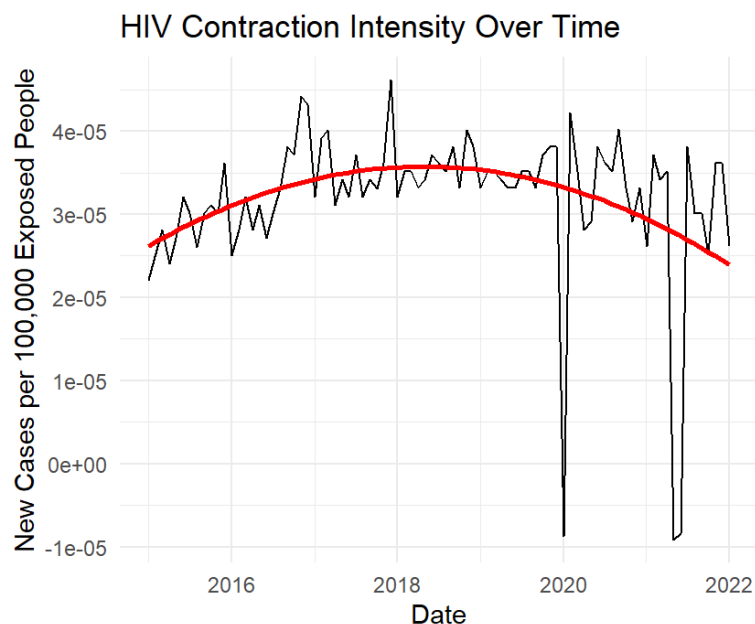


Fig. 6. HIV contraction intensity over time

Another critical extension involves incorporating reverse transitions from AIDS to HIV, enabled by ART. Estimating parameters for $A \rightarrow H$ transition in a 4-state Markov model would more realistically reflect disease dynamics and improve the model's descriptive and predictive accuracy. However, this extension requires access to detailed medical data on patient-level health outcomes and ART-adherence information, which are currently limited.

Finally, improving the model's demographic realism could be achieved by estimating the inflow of newly observed individuals aged 15 and above, which constitutes the age threshold for inclusion in most HIV surveillance programs. Modeling this influx explicitly would support a more robust population-level framework, accommodating both epidemiological transitions and demographic processes.

In summary, the proposed refinements—semi-Markov modeling, inclusion of reverse transitions, and dynamic population entry—represent promising directions for improving the accuracy and applicability of stochastic models of HIV/AIDS in Ukraine. These extensions would allow for more nuanced analysis of epidemic trajectories and provide more informative guidance for public health policy and intervention planning.

In this article, we constructed and justified a reasonable model for the process of HIV/AIDS diagnostics. We fitted continuous-time individual-level and population-level Markov processes to describe the HIV/AIDS epidemic in Ukraine, estimating the discrete-time transition matrix P and its continuous-time counterpart – the generator matrix Q . We managed to retrieve information about distributions lost due to interval censoring, obtained the mean vector of the population Markov process $x(t)$, and its covariance matrix $C(t)$. Finally, we used Monte Carlo simulations to build confidence intervals for the counting process of the total number of people living with HIV/AIDS.

Authors' contribution: Olga Vasylyk - conceptualization and methodology; Yehor Okuniey - methodology and research, software, simulation, and analysis of the results, and preparation of the literature review.

Acknowledgments. The authors would like to thank the anonymous referees for their useful and detailed comments on an earlier version of the manuscript.

Sources of funding. This study did not receive any grant from a funding institution in the public, commercial, or non-commercial sectors.

References

- Barnett, T., Whiteside, A., Khodakevich, L., Kruglov, Y., & Steshenko, V. (2000). The HIV/AIDS epidemic in Ukraine: its potential social and economic impact. *Social science & medicine*, 51(9), 1387–1403. [https://doi.org/10.1016/s0277-9536\(00\)00104-0](https://doi.org/10.1016/s0277-9536(00)00104-0)
- Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2), 187–202. https://doi.org/10.1007/978-1-4612-4380-9_37
- Fowler, M., Flynn, P., & Aizire, J. (2018). What is new in perinatal HIV prevention? *Current opinion in pediatrics*, 30(1), 144–151. <https://doi.org/10.1097/mop.0000000000000579>
- HIV/AIDS.Data. (2025). *Statistics on HIV/AIDS in Ukraine collected by State Enterprise "Center of Public Health" of the Ministry of Health of Ukraine*. <https://phc.org.ua/kontrol-zakhvoryuvan/vilsnid/statistika-z-vilsnidu>
- Jackson, C. (2007). *Multi-state modelling with R: the msm package*. Cambridge. <https://doi.org/10.32614/CRAN.package.msm>
- Johnson, N., Kemp, A., & Kotz, S. (2005). *Univariate discrete distributions*. John Wiley and Sons. <https://doi.org/10.1002/0471715816>
- Kleinman, S., & Busch, M. (2000). The risks of transfusion-transmitted infection: direct estimation and mathematical modelling. *Best Practice and Research Clinical Haematology*, 13(4), 631–649. <https://doi.org/10.1053/beha.2000.0104>
- Kruglov, Y., Kobyshecha, Y., Salyuk, T., Varetska, O., Shakarishvili, A., & Saldanha, V. (2008). The most severe HIV epidemic in Europe: Ukraine's national HIV prevalence estimates for 2007. *Sexually transmitted infections*, 84(1), 37–41. <https://doi.org/10.1136/sti.2008.031195>

- Miranda, M., Pingarilho, M., Pimentel, V., Martins, M. R. O., Vandamme, A.-M., Bobkova, M., Böhm, M., Seguin-Devaux, C., Paredes, R., Rubio, R., Zazzi, M., Incardona, F., & Abecasis, A. (2021). Determinants of hiv-1 late presentation in patients followed in europe. *Pathogens*, 10(7), 835. <https://doi.org/10.3390/pathogens10070835>
- Parczewski, M., Gökengin, D., Sullivan, A., de Amo, J., Cairns, G., Bivol, S., Kuchukhidze, G., Vasylyev, M., & Rockstroh, J. K. (2025). Control of HIV across the WHO European region: progress and remaining challenges. *The Lancet Regional Health–Europe*, 52, 1–14. <https://doi.org/10.1016/j.lanepe.2025.101243>
- Raboud, J., Rae, S., Woods, R., Harris, M., Montaner, J., & INCAS and AVANTI Study Groups. (2002). Consecutive rebounds in plasma viral load are associated with virological failure at 52 weeks among HIV-infected patients. *Aids*, 16(12), 1627–1632. <https://doi.org/10.1097/00002030-200208160-00008>
- State Statistics Service of Ukraine. (2022). *Mortality rates by regions of Ukraine, 2015–2022 [data set]*. https://www.ukrstat.gov.ua/operativ/operativ2022/ds/kjp_reg/arh_kjp_reg2022_u.html

Отримано редакцією журналу / Received: 08.08.25
Прорецензовано / Revised: 27.11.25
Схвалено до друку / Accepted: 16.04.26
Опубліковано / Published: 05.06.26

Ольга ВАСИЛИК, д-р фіз.-мат. наук, доц.

ORCID ID: 0000-0002-0880-3751

e-mail: vasylyk@matan.kpi.ua

Національний технічний університет України “Київський політехнічний інститут імені Ігоря Сікорського”, Київ, Україна

Єгор ОКУНЄВ, магістр

ORCID ID: 0009-0009-3695-490X

e-mail: egorky96@gmail.com

Національний технічний університет України “Київський політехнічний інститут імені Ігоря Сікорського”,
Київська школа економіки, Київ, Україна

МОДЕЛЮВАННЯ ЕПІДЕМІЇ ВІЛ/СНІД В УКРАЇНІ ЯК МАРКОВСЬКИЙ ПРОЦЕС ІЗ НЕПЕРЕРВНИМ ЧАСОМ

Упродовж тривалого часу Україна входить до переліку країн з найвищими темпами епідемії ВІЛ/СНІД. У представленій роботі оцінено параметри марковського ланцюга з неперервним часом, щоб описати процес до діагностування ВІЛ/СНІД та, врешті-решт, смерть індивіда, і в такий спосіб отримати модель епідемії на рівні всієї популяції. Завдяки аналізу здійснено спробу відновити інформацію, втрачену через ліве та інтервальне цензурування, наявне в емпіричних даних. Запропонована методологія поєднує аналіз виживання, оцінювання максимальної правдоподібності, моделювання методом Монте-Карло та побудову довірчих інтервалів. Отримані результати дають змогу краще зрозуміти динаміку епідемії ВІЛ/СНІД в Україні та надалі можуть бути використані для актуарних досліджень епідемії, розглядаючи зараження ВІЛ як страховий ризик.

Ключові слова: моделі виживання, кратні декременти, моделі Маркова, симуляція, випадкові процеси.

Автори заявляють про відсутність конфлікту інтересів. Спонсори не брали участі в розробленні дослідження; у зборі, аналізі чи інтерпретації даних; у написанні рукопису; в рішенні про публікацію результатів.

The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses or interpretation of data; in the writing of the manuscript; in the decision to publish the results.