International Journal of Infectious Diseases 165 (2026) 108495

Contents lists available at ScienceDirect

International Journal of Infectious Diseases

journal homepage: www.elsevier.com/locate/ijid

Short Communication

HIV-1 late diagnosis: Strategies to overcome the misclassification of individuals acutely infected with HIV-1 as individuals diagnosed late

Mafalda N.S. Miranda[1,*], Victor Pimentel[1], André Santos[1], André Alemão[1], Fátima Gonçalves[2], Joaquim Cabanas[2], Inês Costa[2], Isabel Diogo[2], Sandra Fernandes[2], Sofia G. Seabra[1], Perpétua Gomes[2,3], Marta Pingarilho[1,#], Ana Abecasis[1,#], on behalf of the on behalf of the Portuguese HIV-1 Resistance Study Group

[1] *Global Health and Tropical Medicine (GHTM), Associate Laboratory in Translation and Innovation Towards Global Health (LA-REAL), Institute of Hygiene and Tropical Medicine, NOVA University of Lisbon (IHMT/UNL), Lisbon, Portugal*
[2] *Laboratório de Biologia Molecular (LMCBM, SPC, CHLO-HEM), Lisbon, Portugal*
[3] *Centro de Investigação Interdisciplinar Egas Moniz (CiiEM), Instituto Universitário Egas Moniz, Costa da Caparica, Portugal*

ARTICLE INFO

Article history:
Received 17 October 2025
Revised 13 February 2026
Accepted 18 February 2026

Keywords:
HIV-1 infection
Late diagnosis
Ambiguity rate
Estimated time of infection

ABSTRACT

Objectives: Late HIV diagnosis is associated with a higher impact on treatment outcomes and a potential for prolonged transmissibility of HIV-1 infection. The consensus definition for late HIV diagnosis is problematic. It was updated in 2022; however, this definition relies on information that might not be clinically available. This study aimed to assess late HIV diagnosis using alternative parameters, in addition to the definition of clusters of differentiation (CD4) cell count, namely, sequence ambiguity rate and estimated time of infection inferred through phylogenetic analysis.

Methods: Clinical, socio-demographic, and genotypic information from 3668 antiretroviral therapy–naïve individuals living with HIV was retrieved from the REGA database. Individuals were classified according to three approaches: (i) CD4 cell count, (ii) sequence ambiguity rate, and (iii) phylogenetic reconstruction using TreeTime to estimate the time of most recent common ancestor (MRCA) as a proxy for time of infection.

Results: Based on CD4 cell count, 53.8% of individuals had a late diagnosis and 46.2% had a non-late diagnosis. Based on sequence ambiguity rate, 57.8% had a chronic and 42.2% had a recent infection, and 86.4% had an estimated time of infection of more than 3 years, whereas 13.6% had less than 3 years. A total of 114 individuals were classified as diagnosed late by CD4 criteria and showed evidence of recent infection based on low ambiguity rates and MRCA estimates under 3 years. These individuals had significantly lower viral loads than those with true late diagnoses (median 61,358 vs 134,730 copies/ml; $P < 0.001$). Overall, 41% of individuals were consistently classified across all three methods.

Conclusions: The definition of late diagnosis remains a major challenge. Alternative and complementary methods, such as the use of viral loads, combined with some more clinical information, may improve the lack of baseline data.

Introduction

By the end of 2024, 40.8 million people were living with HIV (PWH) and 1.3 million new infections occurred [1]. Late diagnosis presents a global concern because of consistently high prevalence rates worldwide. This results in clinical consequences for affected individuals, as well as on prolonged transmission of HIV-1 from undiagnosed individuals. Some individuals diagnosed late may already present with AIDS symptoms upon diagnosis, increasing the difficulty to achieve optimal immunologic response to antiretroviral drugs [2,3].

In 2010 [4], a consensus definition for late HIV diagnosis was established and later updated in 2022 by Croxford *et al.* [5]. The updated definition suggested that researchers should inherently rely on evidence (or lack of evidence) of recent diagnosis. This updated definition should be based on the evaluation of people with

* Corresponding author.
E-mail address: mafaldansmiranda@ihmt.unl.pt (M.N.S. Miranda).
Shared last author.

https://doi.org/10.1016/j.ijid.2026.108495

evidence of a recent infection, and their reclassification as individuals with non-late diagnosis[4].

Most of the time, some important information used by the updated definition is missing. In those cases, the only available parameter for classification of individuals diagnosed late vs non-late remains to be the clusters of differentiations (CD4) cell count with its inherent bias. Individuals with acute infection can be classified as individuals diagnosed late due to a transient low CD4 cell count nadir upon the initial drop of CD4 soon after the infection. For all the previously mentioned reasons, finding informative and accessible parameters to overcome the classification bias in late diagnosis is of extreme importance to HIV research and public health.

This study aims to explore the usability of different parameters to improve the classification of late HIV diagnosis, as well as of acute HIV cases. To that purpose, we used three different methods: one based on CD4 cell count, one based on sequence ambiguity rate, and the last one based on the time of infection through the reconstruction of phylogenetic trees based on most recent common ancestor (MRCA).

Methods

Study group

Clinical, socio-demographic, and genomic information from 3668 antiretroviral therapy–naïve individuals living with HIV from the REGA database [6].

Subtyping

HIV-1 subtyping was performed using the consensus of the result obtained based on three different subtyping tools: Rega HIV Subtyping Tool version 3.46 [7], COMET [8], and Geno2pheno [9].

Classification: CD4 cell count

For the classification based on CD4 cell count, we classified the individuals as late diagnosis if the CD4 cell count was lower than 350 cells/mm^3 and non-late diagnosis if the CD4 cell count was equal or higher than 350 cells/mm^3.

Classification: ambiguity rate of sequences

We classified individuals according to the following: a recent infection if the ambiguity rate of the sequences was lower than 0.47% and individuals with a chronic infection if the ambiguity rate of the sequences was equal or higher than 0.47% [10].

Classification: reconstruction of phylogenetic trees based on the MRCA

We classified individuals with a recent infection if the estimated time of infection was lower than 3 years and individuals with a chronic infection if the estimated time of infection was equal or higher than 3 years.

Acute cases analysis

To analyze the possible acute cases erroneously classified as late diagnosis, we used all the individuals with the primary definition of late diagnosis according to CD4 cell count, then we converged the other two methods—individuals classified as non-late diagnosis/recent infection—to quantify how many individuals could actually be acute cases.

Table 1
Clinical characteristics of the study population in relation to the definitions of late HIV diagnosis.

Characteristics	Total
3668 individuals	N (100%)
Clusters of differentiations 4 cell count	3668 (100)
Late diagnosis (<350)	1974 (53.8)
Non-late diagnosis (≥350)	1694 (46.2)
Ambiguity rate	3668 (100)
Recent infection (<0.47%)	1548 (42.2)
Chronic infection (≥0.47%)	2120 (57.8)
Estimated time of infection based on most recent common ancestor	3668 (100)
More than 3 years	3168 (86.4)
Less than 3 years	500 (13.6)

Results

Characteristics of the study population

Among the 3668 PWH included in the analysis, 66% were males and the median age at time of drug resistance testing was 38 years old (interquartile range: 31-47). Most individuals were originated from Portugal (68.3%) and 23.6% were originated from the African continent. The main transmission route was through heterosexual contact (41.2%) and the most prevalent subtype was subtype B (38.1%) (ST1).

The clinical characteristics of these individuals based on the definitions used were as follows: 53.8% of individuals were diagnosed late and 46.2% were diagnosed non-late based on CD4 cell count; 57.8% had a chronic infection and 42.2% had a recent infection based on sequence ambiguity rate; 86.4% had an estimated time of infection based on MRCA of more than 3 years, whereas 13.6% had an estimated time of infection MRCA of less than 3 years (Table 1).

Analyzing this for concordance of classification through the three methods, we found a concordance between the CD4 cell count and ambiguity rate of 64% and a concordance of 55.8% between CD4 count and the estimated time of infection. The concordance between ambiguity rate and the estimated time of infection was 62%.

A total of 41% of individuals were consistently classified through the combination of these three methods (Figure 1).

A total of 114 (3.1%) individuals—so called false positives—were classified as individuals diagnosed late based on CD4 cell count. They had lower values of ambiguity rate and less than 3 years of time of infection, according to the phylogenetic tree estimation. Compared with the true positives (individuals with CD4 >350, higher ambiguity, and high estimated time MRCA), these individuals had lower values of viral load (median = 61,358.5 copies/ml, compared with median =134,730 copies/ml, P <0.001).

Discussion

The definition of late HIV-1 diagnosis has been a matter of complex discussion within the scientific community. A well-recognized limitation of CD4-based definitions is the transient and sometimes profound decline in CD4 cell counts that can occur during acute infection, which may lead to the misclassification of people with recent infection as being diagnosed late. Although seemingly easy to solve this problem, addressing this bias has proven challenging in practice, largely because key information required by updated definitions—such as the date of the last negative HIV test or laboratory or clinical evidence of acute infection—is frequently unavailable in routine clinical and surveillance settings.

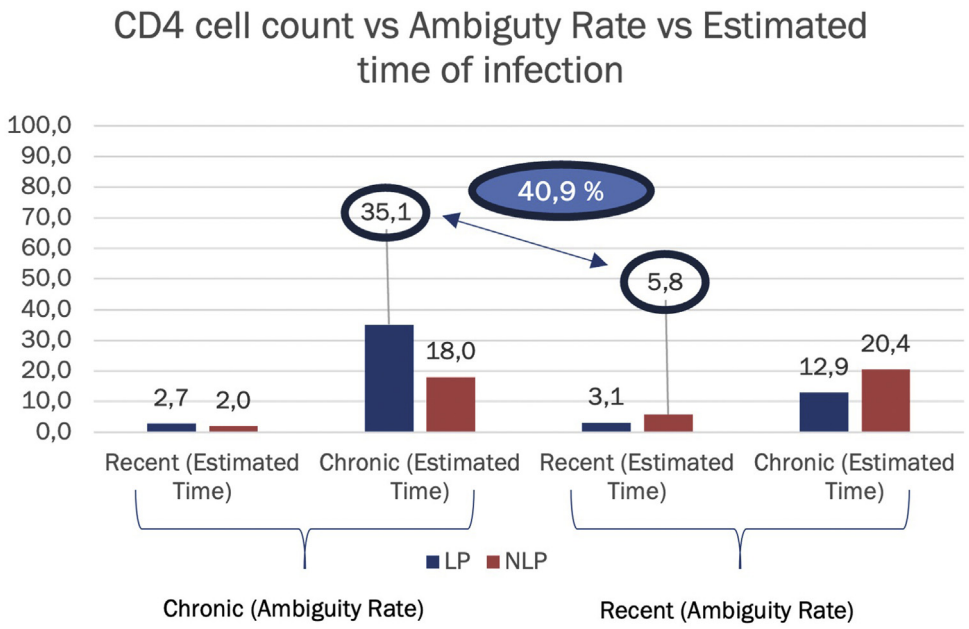

Figure 1. The three combined methods. CD4 cell count (LP in blue; NLP in red) combined with the ambiguity rate of the sequences (left side of the graph chronic infection; right side of the graph recent infection) and with the estimated time of infection (recent and chronic in both sides of the graph). The 35.1% represents the cases correctly identified as late diagnosis (late diagnosis based on CD4 cell count + chronic based on estimated time of infection + chronic infection based on ambiguity rate) and the 5.8% represents the cases correctly identified as non-late diagnosis (non-late diagnosis based on CD4 cell count + recent based on estimated time of infection + recent based on ambiguity rate).
CD, clusters of differentiations; LP, late presentation; NLP, non-late presentation.

In this study, we directly explored this source of misclassification by comparing the classical CD4-based definition of late diagnosis with two alternative indicators of infection recency: sequence ambiguity rates and phylogenetically inferred time to MRCA.

The results found herein highlight the complexity involved in accurately defining acute cases vs late diagnosis among PWH. Although the widely used definition based on CD4 cell counts below 350 cells/mm^3 remains standard, we observed that relying only on this parameter alone leads to misclassification of acute cases as individuals diagnosed late. Generating a known classification bias previously reported in the literature [5].

The incorporation of methods such as sequence ambiguity rates and phylogenetic analysis through time to MRCA proved to be a promising strategy. The observed overall concordance across the three methods is far away to be a perfect indicator of acute cases vs individuals diagnosed late. Although complementary, none of these approaches are entirely overlapping. This study highlights a need for a combined strategy to minimize misclassification when clinical information about recent infection is unavailable. Such a combined strategy will need to be validated based on the training of large data sets of individuals where information about age of infection, CD4 counts, and ambiguities rates is known.

The identification of 114 individuals as false positives—classified as diagnosed late based on CD4 counts—is particularly relevant. These individuals exhibited significantly lower viral loads than individuals truly diagnosed late, supporting the hypothesis that they were likely in an early phase of infection and indicating that viral loads could be an indicator of HIV acute infection when combined with other factors. In this case, the inclusion of viral loads as a complementary marker could be potentially important in future studies, particularly, because it is widely available in clinical settings comparing with CD4 at diagnosis. This phase is characterized by a transient CD4 cell depletion. This finding reinforces that CD4 count alone is not a robust marker for late diagnosis classification, especially in the absence of other clinical or laboratory data relevant for understanding disease progression.

In concordance with previous studies, such as those by Andersson *et al.* [10] and Poljak *et al.* [11], our findings support the usefulness of the sequence ambiguity rate as a marker for recent vs chronic infection. However, the practical application of this parameter can also be biased due to technical constraints. The quality of sequences can vary based on the method used, sequence editing, and length. The phylogenetic analysis using TreeTime seems more accurate in estimating non-late diagnosis. However, the bias in this parameter is also dependent on the robustness of the data and phylogenetic modeling used. Also, this may not be available in many clinical settings, namely, in countries where drug resistance testing is not performed routinely and, therefore, no genomic sequences are available from the individuals.

Taken together, these limitations argue against the use of any single parameter as a standalone solution. Our results also underscore the importance of developing accessible and standardized tools that can be applied on a large scale, particularly, in regions where access to comprehensive diagnostic testing is limited. The combination of a molecular approach with clinical and demographic data can provide a more accurate classification of late diagnosis that will allow to avoid the misclassification of acute cases.

From a public health perspective, improving the accuracy of late diagnosis classification has implications beyond individual patient management. More precise differentiation between acute infection and true late diagnosis can strengthen epidemiologic surveillance, inform the evaluation of testing strategies, and improve the targeting of prevention and linkage-to-care interventions. All in all, such an approach may contribute to improve decision-making in clinical practice, such as resource allocation, intervention strategies, antiretroviral therapy prescription, and better-informed HIV public health policies.

Conclusion

The definition of late diagnosis is still a major challenge in HIV diagnosis. This study explored alternative and complementary

methods to overcome the lack of clinical information at baseline. Combining these techniques with clinical information is crucial for a correct discrimination of individuals truly diagnosed late compared with individuals with acute infection who present with low CD4 values.

Declaration of competing interest

The authors have no competing interests to declare.

Funding

This study was funded by national funds through FCT- Fundação para a Ciência e Tecnologia, I.P., under the R&D unit Global Health and Tropical Medicine (GHTM-UID/004412/2025) and the Associated Laboratory in Translation and Innovation Towards Global Health REAL (LA/P/0117/2020), also by Integriv project (financed by FCT: PTDC/SAUINF/31990/2017) and the MARVEL project (financed by FCT: PTDC/SAU-PUB/4018/2021). This study was financed by the Gilead Génese program through funding to the project HIVLatePresenters.

Author contributions

Conceptualization, M.N.S.M., M.P., and A.A.; methodology, M.N.S.M., M.P., V.P., and A.A.; software, M.N.S.M., V.P., A.S., S.G.S.; validation, M.N.S.M., M.P., and A.A.; formal analysis, M.N.S.M., V.P., A.S., S.G.S., M.P., and A.A.; investigation, M.N.S.M., M.P., V.P.; resources, F.G., J.C., I.C., I.D., S.F., P.G.; writing—original draft preparation, M.N.S.M., M.P., and A.A.; writing—review and editing, M.N.S.M., M.P., and A.A.; visualization, M.N.S.M., M.P., V.P., and A.A.; supervision, A.A.; project administration, A.A.; and funding acquisition, A.A.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ijid.2026.108495.

References

[1] World Health Organization *HIV and AIDS* https://www.who.int/news-room/fact-sheets/detail/hiv-aids.
[2] Guaraldi G, Zona S, Menozzi M, Brothers TD, Carli F, Stentarelli C, et al. Late presentation increases risk and costs of non-infectious comorbidities in people with HIV: an Italian cost impact study. *AIDS Res Ther* 2017;**14**:8. doi:10.1186/s12981-016-0129-4.
[3] Bath RE, Emmett L, Verlander NQ, Reacher M. Risk factors for late HIV diagnosis in the East of England: evidence from national surveillance data and policy implications. *Int J STD AIDS* 2019;**30**:37–44. doi:10.1177/0956462418793327.
[4] Antinori A, Coenen T, Costagiola D, Dedes N, Ellefson M, Gatell J, et al. Late presentation of HIV infection: a consensus definition. *HIV Med* 2011;**12**:61–4. doi:10.1111/j.1468-1293.2010.00857.x.
[5] Croxford S, Stengaard AR, Brännström J, Combs L, Dedes N, Girardi E, et al. Late diagnosis of HIV: an updated consensus definition. *HIV Med* 2022;**23**:1202–8. doi:10.1111/hiv.13425.
[6] Libin P, Beheydt G, Deforche K, Imbrechts S, Ferreira F, Van Laethem K, et al. RegaDB: community-driven data management and analysis for infectious diseases. *Bioinformatics* 2013;**29**:1477–80. doi:10.1093/bioinformatics/btt162.
[7] Pineda-Peña AC, Faria NR, Imbrechts S, Libin P, Abecasis AB, Deforche K, et al. Automated subtyping of HIV-1 genetic sequences for clinical and surveillance purposes: performance evaluation of the new REGA version 3 and seven other tools. *Infect Genet Evol* 2013;**19**:337–48. doi:10.1016/j.meegid.2013.04.032.
[8] Struck D, Lawyer G, Ternes AM, Schmit JC, Bercoff DP. COMET: Adaptive context-based modeling for ultrafast HIV-1 subtype identification. *Nucleic Acids Res* 2014;**42**:1–11. doi:10.1093/nar/gku739.
[9] Pirkl M, Büch J, Friedrich G, Böhm M, Turner D, Degen O, et al. Geno2pheno: recombination detection for HIV-1 and HEV subtypes. *NAR Mol Med* 2024;**1** ugae003. doi:10.1093/narmme/ugae003.
[10] Andersson E, Shao W, Bontell I, Cham F, Cuong DD, Wondwossen A, et al. Evaluation of sequence ambiguities of the HIV-1 pol gene as a method to identify recent HIV-1 infection in transmitted drug resistance surveys. *Infect Genet Evol* 2013;**18**:125–31. doi:10.1016/j.meegid.2013.03.050.
[11] Lunar MM, Židovec Lepej S, Poljak M. Sequence ambiguity determined from routine pol sequencing is a reliable tool for real-time surveillance of HIV incidence trends. *Infect Genet Evol* 2019;**69**:146–52. doi:10.1016/j.meegid.2019.01.015.