Interpreting Diagnostic Histories into HIV Infection Time Estimates: Framework and Online Tool

Eduard Grebe[1#*], Shelley N. Facente[2,3*], Jeremy Bingham[1], Christopher D. Pilcher[4], Andrew Powrie[5], Jarryd Gerber[5], Gareth Priede[5], Trust Chibawara[1], Michael P. Busch[3,4], Gary Murphy[6], Reshma Kassanjee[7] and Alex Welte[1*]

[1]South African Centre for Epidemiological Modelling and Analysis (SACEMA), Stellenbosch University, Stellenbosch, South Africa; [2]Facente Consulting, Richmond, CA, USA; [3]Blood Systems Research Institute, San Francisco, CA, USA; [4]University of California San Francisco, San Francisco, CA, USA; [5]Implicit Design, Cape Town, South Africa; [6]Public Health England, London, United Kingdom; [7]Department of Statistical Sciences, University of Cape Town, Cape Town, South Africa.

[*]These authors contributed equally.

[#]Email: eduardgrebe@sun.ac.za; Address: SACEMA, Stellenbosch University, Private Bag X1, 7602 Matieland, South Africa.

Abstract

It is frequently of epidemiological or clinical interest to estimate the date of HIV infection or time-since-infection of individuals. Yet, for more than 15 years, the only widely-referenced infection dating algorithm using diagnostic test results to estimate time-since-infection has been the 'Fiebig staging' system. This defines a number of stages of early HIV infection through various 'standard' combinations of contemporaneous discordant diagnostic results, using tests of different sensitivity.

To develop a new, more nuanced, infection dating algorithm, we generalised the Fiebig approach to accommodate positive and negative diagnostic results generated on the same *or* different dates, and arbitrary current or future tests – as long as the test sensitivity is known. For this purpose, test sensitivity is conceptualised as the probability that a specimen will produce a positive result, expressed as a function of time since infection. This can be summarised as a median 'diagnostic delay' parameter (together with a measure of inter-subject variability).

The present work outlines the analytical framework for infection date estimation using subject-level diagnostic testing histories, and data on test sensitivity. In the typical case, where there is a negative test result and a positive test result on different days, the formal expression of the likelihood function is an interval with a round-shouldered plateau. We also investigate the impact, on infection time estimates, of subject level correlation of results on different tests; finding that such correlation is largely innocuous even in the case of perfect correlation.

Finally, we introduce here a publicly-available online HIV infection dating tool that implements this estimation method, bringing together 1) curatorship of HIV test performance data, and 2) infection date estimation functionality, to calculate plausible intervals within which infection likely became detectable for each individual. The midpoints of these intervals are interpreted as infection time 'point estimates' and referred to as Estimated Dates of Detectable Infection (EDDIs). This tool, available at https://tools.incidence-estimation.org/idt/, is readily updatable as test technology evolves, given the simple architecture of the system and its nature as an open source project.

Introduction

For pathogenesis studies, diagnostic biomarker evaluation, and surveillance purposes, it is frequently of interest to estimate the HIV infection time (date of infection or time-since-infection) of study subjects. Ideally, a biomarker signature would provide reasonable direct estimates of an individual's time-since-infection, but natural inter-subject variability of pathogenesis and disease progression makes this difficult. This work focuses on the use of readily available, qualitative (i.e. positive/negative) test results to estimate dates of infection.

Most simply, nuanced infection dating applies to subjects who produce a negative test result and also (usually at a later time) a positive test result, taking into account that no test can detect infection immediately after acquisition. Hence, infection can at best be estimated to have occurred during an interval in the past, relative to the date(s) of the test(s).

When a subject obtains 'discordant results', i.e. a negative and a positive test result on the same day, this typically manifests as positive results on 'more sensitive' tests than those on which the negative results were obtained. What we mean here by higher sensitivity is a shorter *'typical delay between acquisition and detectability of infection*.' For high-performing diagnostic tests, such as are normal for HIV and other viral infections like hepatitis, test sensitivity is really a function of time-since-infection, which can sometimes be summarised as the probability of correctly identifying a positive case. We further discuss this idea in some detail in a related analysis of the 'residual risk' of infection from blood products that remains despite extensive screening (Welte *et al*., forthcoming).

For more than 15 years, the only widely-referenced infection dating algorithm using test results to estimate time-since-infection has been the 'Fiebig staging' system [1]. This system defines a number of stages of early HIV infection through various 'standard' combinations of contemporaneous discordant results using diagnostic tests of different sensitivity. For example, Fiebig stage 1 is defined as exhibiting reactivity on a viral load assay, but not (yet) on a p24 antigen assay, and in the seminal 2003 paper was estimated to begin approximately 11 days after infection, with a mean duration of 5.0 days [1]. The particular tests used in these original calculations are largely no longer in use, nor commercially available. Others have used newer diagnostic assays to recalibrate the Fiebig stage mean duration estimates or redefine similar stages as an analogue to the Fiebig method [2, 3], though as tests evolve and proliferate, it becomes infeasible to calibrate all permutations of test discordancy.

Infection date estimates will usually be summarised as intervals, the midpoint of which is naturally considered a 'point estimate' of the date of infection. These intervals can be understood as plateaus on a probability curve (Bayesian 'posterior'). The intervals can further serve as 'priors' in the analysis of additional quantitative markers obtained within a calibrated dynamic range.

Building from the Fiebig staging concept, we developed a new, more nuanced infection dating algorithm to meet the needs of a substantial collaboration (the Consortium for the Evaluation and Performance of HIV Incidence Assays – CEPHIA) in support of the discovery, development and evaluation of recent infection biomarkers [4, 5]. The primary CEPHIA activity has been to develop various case definitions for 'recent HIV infection', with intended applicability mainly to HIV incidence surveillance, rather than individual-level staging, although the latter application has also been explored [4, 6]. A key challenge is that, though based on large runs of many specimens under consistent conditions, the primary recency test results bring together material from subjects in numerous studies, each of which used different diagnostic algorithms to capture information about the timing of HIV acquisition or seroconversion. To meet this challenge we linked specimens from thousands of study-patient interactions into a coherent and consistent infection dating scheme (as described in Section 4.3 of Kassanjee, 2014 [7]), in order to estimate the critical properties of recent infection tests for surveillance applications, i.e. Mean Duration of Recent Infection (MDRI) and False Recent Rate (FRR) as previously defined [8].

In order to align diagnostic testing information across multiple sources, one needs a common reference event in a patient history – ideally, the time of infectious exposure. For most of the present exposition, we will talk as if we are directly estimating time of infection. When dealing with actual patient data, and curating actual test properties for our online tool, we explicitly confront the point that in reality we are usually constrained to estimate the time when a particular test (X) would have first detected the infection. We will call this the X-test-specific *Date of Detectable Infection*, or DDI_X.

The present work outlines the analytical framework for infection date estimation using 'diagnostic testing histories', including both extensions and simplifications of the previously described analysis [7]. We also introduce a publicly-available online HIV infection dating tool that implements this estimation, bringing together 1) curatorship of HIV test performance data, and 2) infection date estimation functionality. It is readily updatable as test technology evolves, given the simple general architecture of the system and its nature as an open source project.

Generalised Fiebig-like Staging

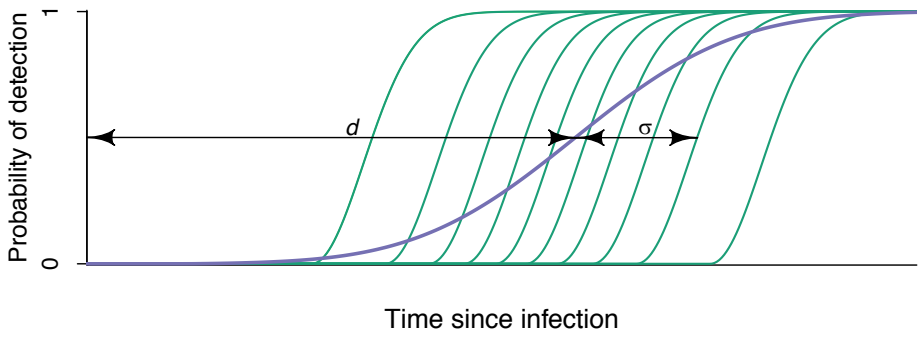
The fundamental feature of the Fiebig staging system [1] is that it identifies a naturally-occurring sequence of discordant diagnostic tests that together indicate early clinical disease progression, where approximate duration of infection can be categorized into stages through analysis of the combination of specific assay results. One can treat these numbered, sequential stages as an arbitrary pathway that all infected persons pass through; however, recalibration of the duration of these stages is extremely difficult given changes in test technology and usage, unless the specific assays informing the estimation are well understood.

As we demonstrate below, it is most robust to interpret any combination of diagnostic test results into an estimated duration of infection, if these tests have been independently benchmarked for diagnostic sensitivity (i.e. a median or mean duration of time from infection to detectability on that assay has been estimated). This more nuanced method allows both for incorporation of results from any available test, and from results of tests run on specimens taken on different days.

Interpreted at a population-level, a particular test's sensitivity curve expresses the probability that a specimen obtained at some time t after infection will produce a positive result. The key features of a test's sensitivity curve (represented by the purple curve in Figure 1) are that:

- there is effectively no chance of detecting an infection immediately after exposure;
- after some time, the test will almost certainly detect an infection;
- there is a characteristic time range over which this function transitions from close to zero to close to one. This can be summarised as something very much like a mean or median and a standard deviation.

Figure 1



Genotype and other attributes, such as concurrent infections, age, the particular invading quasispecies, post-infection factors, etc. affect the performance of a test for a particular individual; this in principle determines a subject-specific curve, such as one of the green curves in Figure 1, which capture the probability, as a function of time, that specimens from a particular subject will produce a positive diagnostic result. Because assay results are themselves imperfectly reproducible even on the same individual, these green curves do not transition step-like from zero to one, but have some finite window of time over which they transition from close to zero to close to one.

The green curves are shifted copies of the same underlying *shape* and *scale*. The *shift* parameter captures something like an individual-level 'diagnostic delay', while the *shape* and *scale* parameters together encapsulate 1) the growth rate of the underlying assay response variable *in the region of the positive/negative threshold*, and 2) the measurement noise *in the region of the diagnostic threshold*. The purple population-level curve, on the other hand, captures the population-level distribution of the individual diagnostic delays – that is, the distribution of 'scale parameters' for the individual (green) curves. The (qualitative) shape of this curve has no particular relationship to the shapes of the individual sensitivity curves – indeed the population curve is not from the same formal family as the individual curves.

In contrast to the usual statistical definition of 'sensitivity' as the proportion of 'true positive' specimens that produce a positive result, we propose to summarise the population-level sensitivity of any particular diagnostic test into one or two 'diagnostic delay' parameters (d and σ in Figure 1). By far the most important parameter is an estimate of '*median diagnostic delay.'* In Figure 1, this is the parameter d. If there were perfect test result conversion for all subjects (i.e. no assay 'noise'), and further no inter-subject variability, this would reduce the smoothly varying purple curve to a step function.

It is not feasible to conduct studies large enough to obtain much detail about the structure of sensitivity curves for a great number of diagnostic tests. Hence, we propose that the distribution of diagnostic delays be approximated by choosing a generic functional form which has roughly the sigmoidal structure seen in studies (our infection dating tool implements a simple cumulative normal distribution), and setting one additional parameter to capture the variability in the diagnostic delay distribution – parameter σ in Figure 1.

To estimate individual infection times, then, one needs to obtain estimates of the median diagnostic delays for all tests occurring in a data set, and then interpret each individual assay result as excluding some 'non-possible' segment of time, ultimately resulting in a final inferred interval of time during which infection likely occurred. The prototypical situations in which one can perform dating, within this paradigm, are then when a subject:

1. tests positive at a study visit after testing negative at a previous study visit, or
2. tests positive on some component of an algorithm, and negative on another component, at a single study visit.

These calculations require that each individual has at least one negative test result and at least one positive test result. In the primitive case where there is precisely one of each, namely a negative result on a test with an expected diagnostic delay of d_1 at t_1 and a positive result on a test with an expected diagnostic delay of d_2 at t_2, then the interval is simply from $(t_1 - d_1)$ to $(t_2 - d_2)$. When there are multiple negative results on tests at $t_i^{(-)}$ each with a diagnostic delay $d_i^{(-)}$, and/or multiple positive results on tests at $t_j^{(+)}$ each with a diagnostic delay $d_j^{(+)}$, then each individual negative/positive test result provides a candidate earliest plausible and latest plausible date of infection. The most informative tests, then, are the ones that functionally narrow the 'infection window' (i.e. result in the latest start and earliest end of the window) by excluding periods of time during which infection would

not plausibly have become detectable. In this case, the point of first 'detectability' refers to the time when the probability of infection being detected by an assay first exceeds 0.5.

This interval will not be any particular confidence interval, as the details of inter-test interval and inter-subject variability matter. When the most informative negative and positive tests are at different timepoints, the estimated date of infection is typically the centre of a very broadly plateaued (rather than 'peaked') likelihood function. Given a broadly uniform prior, this can be interpreted into a round-edged plateau-like Bayesian posterior. Such a posterior, derived from an individual's diagnostic testing history, could also serve as a prior for further analysis, if there is a quantitative biomarker available, for which there is a robustly calibrated maturation/growth curve model. We do not deal with this in the present work, but it is the subject of ongoing analysis (Pilcher *et al*., forthcoming), and an important potential application of this framework and tool.

Interpreting each element of a diagnostic testing history as independently excluding some period of time from the plausible infection time interval frees the analysis from the constraints of a pre-enumerated list of infection stages dependent upon defined assay combinations. It does, however, require estimation of the diagnostic delay for each assay, either by sourcing direct estimates of the diagnostic delay, or by sourcing such data for a biochemically equivalent assay. Our online HIV infection dating tool is preloaded with sensible diagnostic delay estimates for over 60 HIV assays, and users can both add new tests and provide alternative diagnostic delay estimates for those tests which are already loaded.

This interpretation of individual test results superficially appears to rely on the assumption that test results are independent (i.e. uncorrelated). Of course, the very factors that influence the individual sensitivity curves in Figure 1 suggest that strong correlations between results of different tests on the same person are likely. Given this, we provide below a more precise discussion of a formal inference scheme. This discussion demonstrates explicitly 1) when and how test correlation might influence the analysis, and 2) how analysis of qualitative diagnostic testing data can interface with analysis of quantitative markers of infection stage.

Formal Likelihood Function

It is analytically useful to specify an explicit 'likelihood function', i.e. a formula for capturing the probability of seeing a data element (or set), given some hypothetical values for parameters which determine the behaviour of the underlying system, including the measurement process. This facilitates all the usual statistical manipulations for obtaining confidence intervals, Bayesian posteriors, etc. For the present application, test sensitivity curves such as those in Figure 1 are precisely the likelihood of obtaining a positive result upon application of a given test, at a given time since infection. The likelihood of obtaining a negative result, on this very application of the test, is simply 1 minus the likelihood of obtaining a positive result, i.e. a vertically flipped version of the test sensitivity curve. As noted above, meaningful infection dating relies on having at least one negative test result and at least one positive test result.

Classical test conversion series

To begin, we consider precisely one negative and one positive test result, arising from two subject-study interactions, at times t_1 and t_2 respectively, separated by some duration δ. In order to make inferences about the time of infection, we construct a likelihood function which expresses the probability of seeing these two particular results, *as a function of a hypothetical infection time*. This kind of likelihood (of two observations) is usually written as the product of:

- the likelihood of seeing one result (chosen arbitrarily to be considered first) given the hypothetical time of infection, and

- the likelihood of seeing the other result, given
    - the same hypothetical time of infection, and
    - the fact that the other result has in fact been obtained.

Using T_{inf} to denote the actual time of infection that we are trying to estimate, t_{inf} to denote particular values of hypothetical infection time and $[+, t_n]$ and $[-, t_n]$ to denote positive and negative test results at observation times t_n, respectively, this can be written as:

$$\begin{aligned} L(t) \equiv\ & L\big([-, t_1], [+, t_2] \big| T_{inf} = t\big) \\ =\ & L\big([-, t_1] \big| T_{inf} = t\big) \cdot L\big([+, t_2] \big| \{[-, t_1], T_{inf} = t\}\big) \\ =\ & L\big([+, t_2] \big| T_{inf} = t\big) \cdot L\big([-, t_1] \big| \{[+, t_2], T_{inf} = t\}\big) \end{aligned}$$

capturing that the likelihood of seeing both of two events (A and B) is equal to either

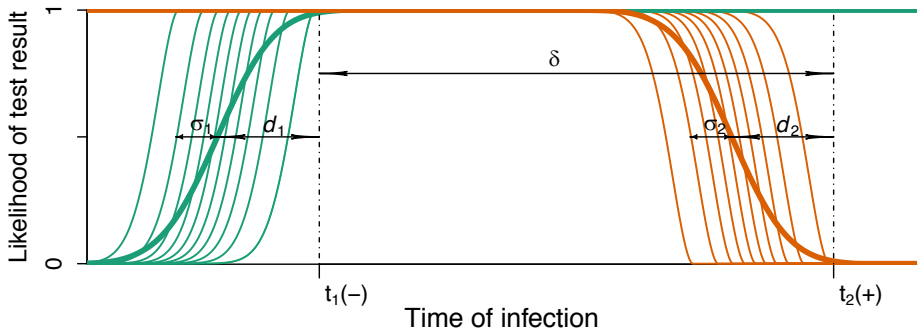
1. *the likelihood of A* multiplied by *the likelihood of B, given A*, i.e. $L(A) \cdot L(B|A)$, or

2. *the likelihood of B* multiplied by *the likelihood of A, given B*, i.e. $L(B) \cdot L(A|B)$.

The details of the *conditioned* likelihoods, which might be complex, must necessarily be such that the two formulations are equivalent. We will focus in detail on the first formulation, as it seems more intuitively appealing when $t_1 < t_2$.

Figure 2 shows, in thick green and red, respectively, the *population-level* likelihoods of observing the negative test result at t_1 and the positive test result at t_2. A subset of the family of *individual-level* curves, chosen to visually suggest their distribution, is indicated as thin lines. A close look at these curves reveals that they are the horizontally flipped (and in the case of the green curves, also vertically flipped) test sensitivity curves of the tests performed (compare with the detailed shapes in Figure 1). These curves display information for each test result, considered independently.

Figure 2



The fundamental point of estimating an infection time is that both tests were in fact performed on the same individual. It is highly likely that those individuals who convert rapidly, post infection, on $Test_1$ also convert rapidly on $Test_2$ – which might, after all, be the same test, and is likely to be a similar test. The details of this conditioning can in principle be complex, and it is infeasible to study all the correlations between all tests in use in studies. A critical question, then, is whether, when, and how this correlation impacts the conditioned likelihoods which are the fundamental building block of a forma inference of infection time from diagnostic testing histories.

The 'worst case' scenario would be when the correlation is very strong, as it would be if the tests performed at the two times are in fact the same test. We have explicitly implemented a model of test sensitivity based on the following points:

- the performance of any test is defined by a family of N individual-level sensitivity curves of the type in Figures 1 and 2.

- for a particular test, each individual-level curve is a shifted Weibull with the same shape and scale parameter.

- The shift parameter is normally distributed, though with a discretised realisation, with step $\epsilon = \frac{1}{N}$, i.e. we assign individual diagnostic delays (Weibull shift parameters) to the percentiles $\frac{1}{2}\epsilon, \frac{3}{2}\epsilon, \dots, 1 - \frac{1}{2}\epsilon$ of a normal distribution.

- The mean of the distribution of shift parameters is a test's mean diagnostic delay (d in Figures 1 and 2), and the standard deviation (σ in Figures 1 and 2) manifests as something akin to a shape parameter of the population-level curve.

To keep the scenario simple initially, we first consider the case when the two test times differ by more than $d + \sigma$. We later consider the complication of the other extreme, i.e. when the positive and negative test results are obtained on the same day, and the distributions of the diagnostic delays overlap substantially.

The behaviour of the fully-conditioned likelihood expression

$$L\big([-, t_1], [+, t_2] \big| T_{inf} = t\big) = L\big([-, t_1] \big| T_{inf} = t\big) \cdot L\big([+, t_2] \big| \{[-, t_1], T_{inf} = t\}\big)$$

can then be understood by considering how the factor $L\big([+, t_2] \big| \{[-, t_1], T_{inf} = t\}\big)$ might differ from the naïve population-averaged, unconditioned $L\big([+, t_2] \big| T_{inf} = t\big)$. The latter is what one can obtain from a study investigating the performance of one or several diagnostic tests, without having to apply the particular test combinations to particular individuals.

We now analyse the various ranges of t_{inf} which are qualitatively different from each other:

Values of t_{inf} on the far-left end of the timeline: For very 'early' hypothetical infection times, the likelihood of seeing the negative result at t_1 becomes very small. If that negative result has indeed occurred at t_1, it would normally be the result of a laboratory error, which would have a reasonable chance of being detected with strong quality controls. If the error remains undetected, testing positive at t_2 (a time later than t_1 by a significant margin) is nevertheless almost assured, i.e.

$$L\big([+, t_2] \big| \{[-, t_1], T_{inf} = t\}\big) \approx L\big([+, t_2] \big| T_{inf} = t\big) \approx 1$$

So, in this case there is no discernible difference between the unconditioned and conditioned likelihood, although there is no analytical cure for a false negative result.

***Values of t_{inf} within the dynamic range of the Test$_1$ sensitivity curve*:** Figures 3a-3e consider a series of hypothetical infection times (t_{inf}) that span the likely range of diagnostic delays.

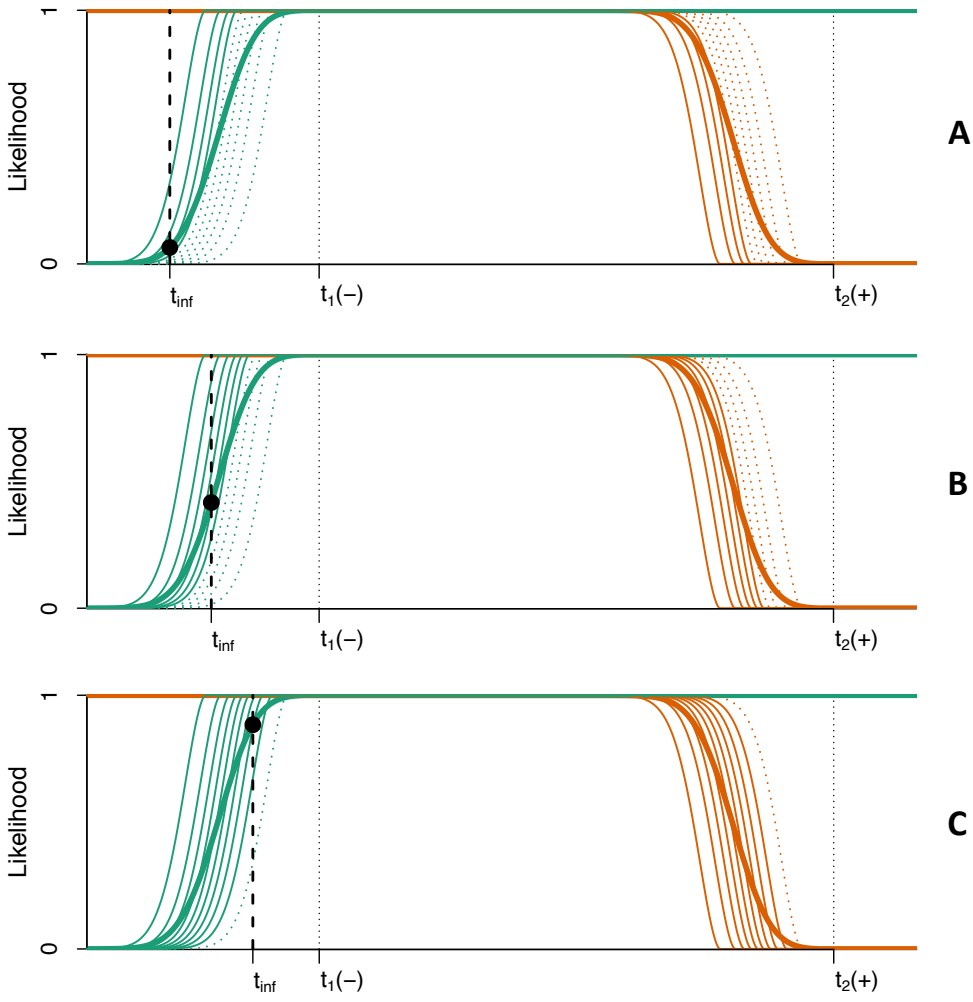
Figure 3a, indicating that the negative result at t_1 occurs somewhat longer after infection than the mean diagnostic delay, is suggestive of the subject being a significantly slower-than-average progressor on the diagnostic marker. This is captured by the dotted green (faster) individual progression curves for Test$_1$, indicating their reduced plausibility. Correspondingly, only the slowest progression rates are plausible among the red curves for Test$_2$. Nevertheless, given the location of t_{inf}, namely long before the application of Test$_2$, it does not matter which of the Test$_2$ progression curves the individual is likely to be on – they all evaluate to 1 so long after infection.

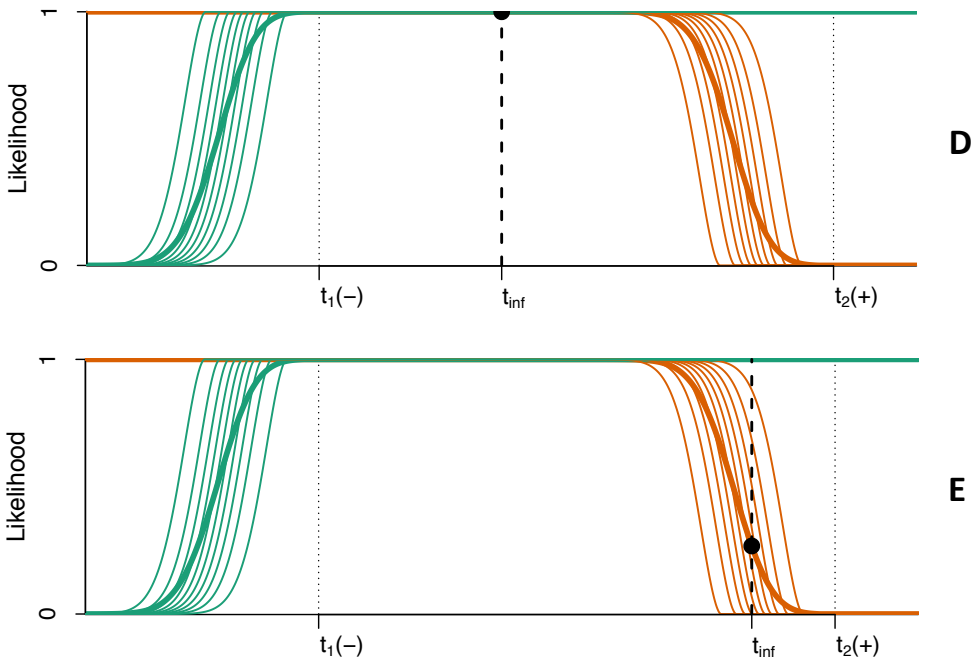
Figure 3b, indicating that the negative result at t_1 occurs at a time after infection approximately equal to the mean diagnostic delay, is suggestive of the subject not being a significantly faster-than-average progressor on the diagnostic marker. This is captured by the reduced number of dotted green (fastest) individual progression curves for Test_1. Correspondingly, the fastest progression rates are less plausible among the red curves for Test_2. Once more, given the location of the hypothetical infection time, namely long before the application of Test_2, it does not matter which of the Test_2 progression curves the individual is likely to be on – they all evaluate to 1 so long after infection.

Figure 3c, indicating that the negative result at t_1 occurs at a time after infection that is significantly less than the mean diagnostic delay, is consistent with all but one of the green and hence red individual progression lines are plausible. Not only does the negative result at t_1 not imply significant conditioning on the subject's diagnostic marker progression rate, but all the individual-level red curves in any case evaluate to 1 *at the time of* Test_2.

***Values of t_{inf} anywhere near, or to the right of t_1*:** For these 'later' hypothetical infection times, we *expect* to see a negative result for the test at t_1, even more so than in Figure 3c, and so, the negative result provides no information on the question of whether the subject is prone to rapid or slow test conversion. Hence, no modification is implied of $L\left([+,t_2]\middle|\{[-,t_1],T_{inf}=t\}\right)$ relative to the population average $L\left([+,t_2]\middle|T_{inf}=t\right)$, though of course in this region there are many values of t_{inf} for which this likelihood is not approximately 1. Figures 3d and 3e show values of t_{inf} on the 'plateau' and on the 'descent' from the plateau in the dynamic range of diagnostic delays of Test_2.

Figure 3





The three zones of t_{inf} discussed above account for the full range of values of t for which the joint likelihood is to be constructed. It is clear that the full joint likelihood is indeed given by the product of the unconditioned population-level likelihoods for the two test results, as shown in Figure 4. As the curves obtain values indistinguishable from either 0 or 1 for much of their range, this product is little more than a superposition of the two curves.

Figure 4

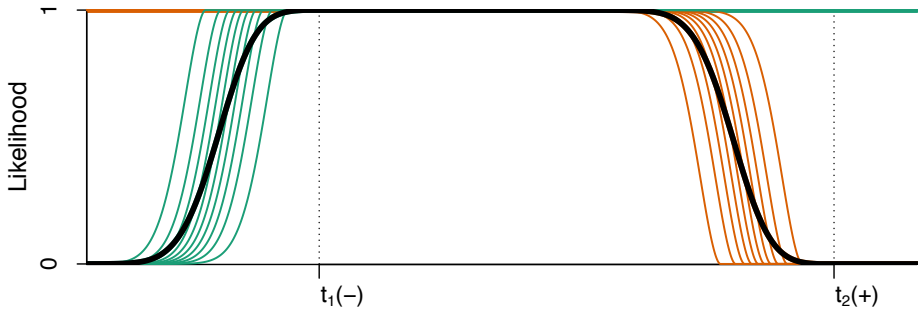
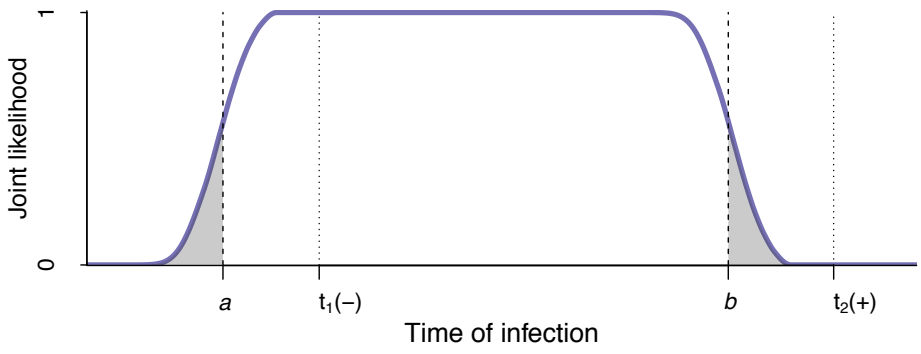


Figure 5 shows where this previously noted round-shouldered plateau is located relative to the test dates and population-averaged diagnostic delays.

Figure 5



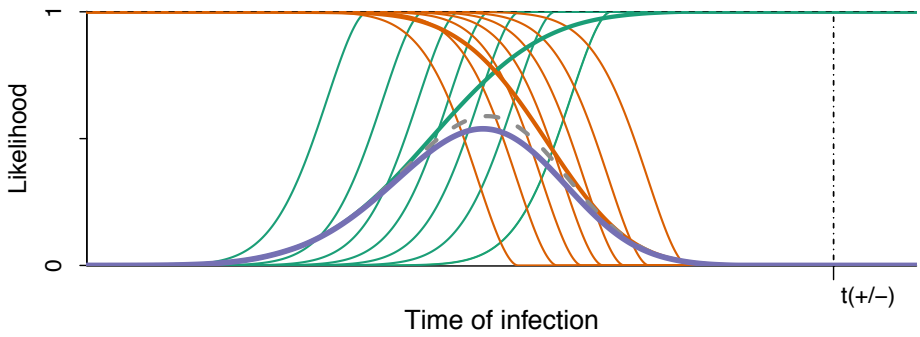
As indicated in Figure 5, under the parametric assumptions outlined above, one may specify a 'confidence level' (such as usually encapsulated in a significance level α, chosen to be 0.05 in the figure) and calculate the bounds of the (in our case, 95%) 'credibility interval' $[a, b]$, encompassing the relevant proportion of the posterior probability density $p(t)$, i.e. we find the values of a and b which satisfy

$$\int_{-\infty}^{a} p(t)\, dt = \int_{b}^{\infty} p(t)\, dt = \frac{\alpha}{2}$$

Discordant results on a given study-visit

Figure 6 shows the typical 'discordant test' situation, where a test with a longer diagnostic delay produces a negative result and a test with a shorter diagnostic delay produces a positive result, at the same visit.

Figure 6



Even here, though not as starkly as in the case where the two tests are conducted at significantly different times, conditioning one result on the other has relatively modest impact. Moving the hypothetical infection time to the left, the negative result becomes less likely, and the effect of the conditioning on the likelihood of seeing the second test result becomes more significant. However, as the hypothetical infection time moves further left, the times under consideration leave the dynamic range of the positive test, and it becomes ever less plausible that a negative test result is obtained. We do not explicitly display figures indicating the conditioning implied for various hypothetical values of infection time, but merely indicate in the solid blue curve the formally calculated fully specified joint likelihood which takes this conditioning into account in terms of the extreme correlation model outlined above. This exact likelihood does not differ meaningfully from the simple product of the population-level likelihoods of the two tests (shown dashed, in grey). The main conclusion, then, is that relative to the test date, plausible infection times are largely located between the two diagnostic delays (with some spreading due to variability).

Figure 7 shows the situation where the dynamic ranges of the tests are essentially the same. In this case, the plausible dates of infection are centred around the shared diagnostic delay of the tests, again with some spread for variability. The relatively small amplitude of the exact curve indicates that the fully conditioned discordancy is significantly less likely to occur than one would infer from a naïve calculation, but the key point is that the infection time estimate is not affected at the level at which it can be plausibly reported.

Figure 7

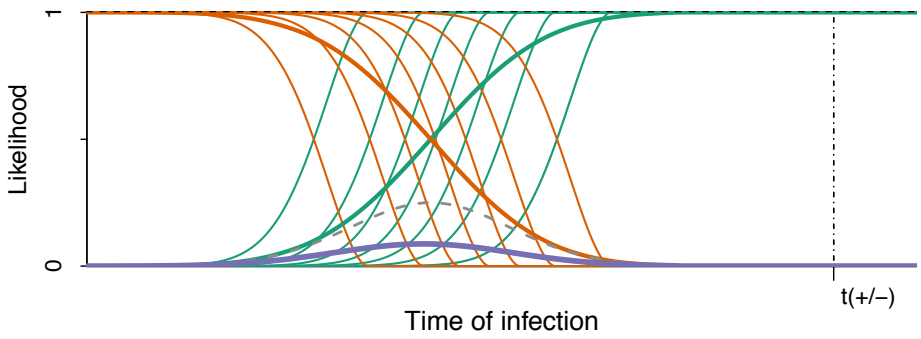
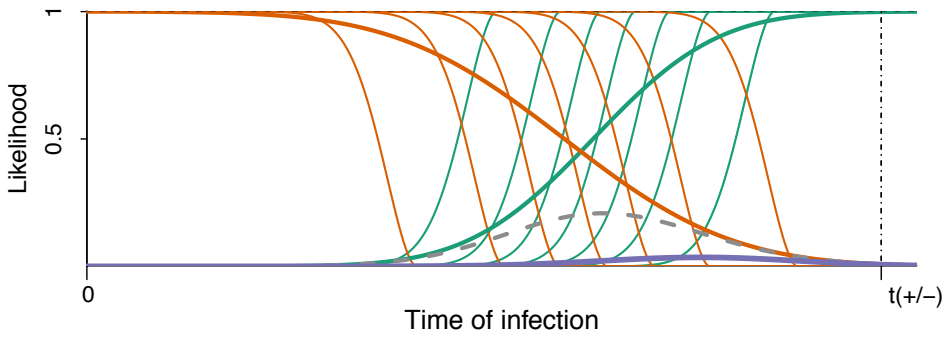


Figure 8 shows an outlier situation in which a more sensitive test is negative while a less sensitive test is positive. Relative to the naïve product of likelihoods, the correctly specified joint likelihood is very small for all values of *t*. This indicates that such anomalous discordant results are extremely rare, arising most plausibly from test error. If such an outlier occurs without test error, the fully-conditioned likelihood could differ significantly from the naïve one; however, it would depend on essentially unknowable details of distributional tails and test correlation, and such cases are sufficiently rare to have no impact on conclusions drawn from observing large numbers of individuals. Note that the extreme rarity of anomalous discordant results is a function of the very strong intra-test correlation assumed in this model; in reality the intra-test correlation is likely far less strong, making these events less rare but also lessening the discrepancy between the naïve and fully-conditioned likelihood.

Figure 8



The Tool

The method outlined above has been implemented in a public online tool, available at https://tools.incidence-estimation.org/idt/. The source code for the tool is available publicly under an open source licence at https://github.com/SACEMA/infection-dating-tool/.

Having established that test correlation is largely irrelevant for the estimation of plausible infection windows and infection time point estimates, we can utilise independently estimated test sensitivities ('diagnostic delays') to interpret individual-level diagnostic test history data into plausible infection time intervals and infection time point estimates for each subject.

In practice, the timing of infectious exposure is seldom known, even in intensive studies, and studies of diagnostic test performance therefore provide *relative* times of test conversion [9-11]. Diagnostic delay estimates are therefore anchored to a standard reference event – the first time that a highly-sensitive viral load assay with a detection threshold of 1 RNA copy/ml of plasma – would detect an infection. We call this the *Date of Detectable Infection* (DDI). The tool endows study subjects with a point estimate of this date, which we call the *Estimated Date of Detectable Infection* (EDDI). The time

from infectious exposure to DDI is likely to be variable between individuals, but the tool does not rely on any assumptions about the average duration of this pre-DDI state.

The key features of our online tool for HIV infection date estimation are that:

1. Users access the tool through a website where they can register and maintain a profile which saves their work, making future calculations more efficient.
2. Individual test dates and positive/negative results, i.e. individual-level 'testing histories', not just algorithm-level diagnoses, can be uploaded in a single comma-delimited text file for a group of study subjects.
3. Estimates of the relative 'diagnostic delay' between the assays used and the reference viral load assay must be provided, with the option of using a curated database of test properties which provides cited estimates for over 60 HIV assays.
    a. If a viral load assay's detection threshold is known, this can be converted into a diagnostic delay estimate via the exponential growth curve model (Pilcher, et al., forthcoming). We assume that after the viral load reaches 1 RNA copy/ml, viral load increases exponentially during the initial ramp-up phase. The growth rate has been estimated at 0.35 $\log_{10}$ RNA copies/ml per day (i.e., a doubling time of slightly less than one day) [1]. The growth rate parameter defaults to this value, but users can supply an alternative estimate.
4. Using the date arithmetic described above, when there is at least one negative test result and at least one positive test result for a subject, the uploaded diagnostic history results in:
    a. a point estimate for the date of first detectability of infection (the EDDI);
    b. an earliest plausible and latest plausible date of detectable infection (EP-DDI and LP-DDI); and
    c. the number of days between the EP-DDI and LP-DDI (i.e., the size of the 'DDI interval'), which gives the user a sense of the precision of the estimate.

The logic and diagnostic test performance data required for infection dating has significant overlap with that required to calculate the residual risk of infectious material being missed by screening algorithms applied to blood products. Therefore, the online tool has a residual risk calculator built into it as well. As the question of residual risk involves additional concepts which deserve proper treatment, this aspect of the tool is discussed and presented in a separate article (Welte et al., forthcoming).

Access / User profiles

Anyone can register as a user of the tool. The tool saves users' data files, and their choices about which diagnostic delay estimates to use for each assay, both of which are only accessible to the user who uploaded them. No person-identifying information is used or stored within the tool; hence, unless the subject identifiers being used to link diagnostic results can themselves be linked to people (which should be ruled out by pre-processing before upload) there is no sensitive information being stored on the system.

Uploading diagnostic testing histories

A single data file would be expected to contain a 'batch' of multiple subjects' diagnostic testing histories. Conceptually, this is a table like the fictitious example in Table 1, which records that:

- one subject (John) was seen on 10 January 2017, at which point he had a detectable vial load on an unspecified qualitative viral load assay, and a negative Western blot result (discordant tests); and

- another subject (Jill) was screened negative on a point-of-care (PoC) rapid test (RT) on 13 September 2016, and then, on 4 February 2017, was confirmed positive by Western blot, having also tested positive that day on the PoC RT.

Table 1

Subject	Date	Test	Result
John	2017-01-10	Qualitative VL	Positive
John	2017-01-10	Western blot	Negative
Jill	2016-09-13	POC RT	Negative
Jill	2017-02-04	POC RT	Positive
Jill	2017-02-04	Western blot	Positive

In order to facilitate automated processing, the tool demands a list of column names as the first row in any input file. While extraneous columns are allowed without producing an error, there must be columns named *Subject, Date, Test* and *Result* (not case sensitive). Data in the subject column is expected to be an arbitrary string that uniquely identifies each subject. Dates must be in the standard ISO format (YYYY-MM-DD).

It is fundamental to the simplicity of the algorithm that assay results be either 'positive' or 'negative'. There are a small number of tests, notably Western blot and Geenius, which sometimes produce 'indeterminate' results (partially, but not fully, developed band structure). We now briefly reconsider Table 1 by adding the minor twist that the Western blot on subject Jill be reported as indeterminate. In this case, the data must be recorded as results on either one or both of two separate tests:

1. a 'Test-Indeterminate' version of the test – which notes whether a subject will be classified either as negative, or 'at least' as indeterminate; and
2. a 'Test-Full' version of the test, which determines whether a subject is fully positive or not.

There is then no longer any use for an un-suffixed version of the original test. The data from table one, with the minor change of an indeterminate result for subject Jill, is shown in Table 2, with differences highlighted. Note that even while John's test results have not changed, his testing history now looks different, as completely negative results are reported as being negative even for the condition of being indeterminate. Jill's indeterminate result on 4 February requires two rows to record, one to report that the test result is not fully negative (positive on 'Western blot Indeterminate'), and one to report that the result is not fully positive (negative on 'Western blot Full'). Once diagnostic delays are provided for these two sub-tests, the calculation of infection dates can proceed without any further data manipulation on the part of the user.

Table 2

Subject	Date	Test	Result
John	2017-01-10	Qualitative VL	Positive
John	2017-01-10	Western blot Indeterminate	Negative
Jill	2016-09-13	POC RT	Negative
Jill	2017-02-04	POC RT	Positive
Jill	2017-02-04	Western blot Indeterminate	Positive
Jill	2017-02-04	Western blot Full	Negative

Provision of test diagnostic delay estimates

As described above, tests are summarised by their diagnostic delays. The database supports multiple diagnostic delay estimates for any test, acknowledging that these estimates may be provisional and/or disputed. The basic details identifying a test (i.e. name, test type) are recorded in a 'tests' table, and

the diagnostic delay estimates are entered as records in a 'test-properties' table, which then naturally allows multiple estimates by allowing multiple rows which 'link' to a single entry in the tests table. A test property entry captures the critical parameter of the 'average' (usually median) diagnostic delay obtained from experimental data and, when available, a measure of the variability of the diagnostic delay (denoted σ).

The system's user interface always ensures that for each user profile, there is exactly one test property estimate, chosen by the user, as 'in use' for infection dating calculations at any point in time. Users need to 'map' the codes occurring in their data files (i.e. the strings in the 'Test' column of uploaded data files) to the tests and diagnostic delay estimates in the database, with the option of adding entirely new tests to the database, which will only be visible to the user who uploaded them. The tool developers welcome additional test estimates submitted for inclusion in the system-default tests/estimates.

Execution of infection dating estimation

The command button 'process' becomes available when an uploaded testing history has no unmapped test codes. Pressing the button leads to values, per subject, for EP-DDI, LP-DDI, EDDI, and DDI interval, which can be previewed on-screen and downloaded as a comma-delimited file.

By default, the system employs simply the 'average' diagnostic delay parameter, in effect placing the EP-DDI and LP-DDI bounds on the DDI interval where the underlying sensitivity curve evaluates to a probability of detection of 0.5. When the size of the inter-test interval (δ) is greater than about 20 times the diagnostic delay standard deviation (σ), this encompasses more than 95% of the posterior probability.

As an additional option, when values for both d and σ are available, and under the parametric assumptions outlined earlier, users may specify a significance level (α), and the system will calculate the bounds of a corresponding credibility interval. The bounds of the central 95% (in the case of $\alpha = 0.05$) of the posterior are labelled the EP-DDI and LP-DDI.

Source code, distribution and modification

The whole code base for the tool is available in a public source code repository (at https://github.com/SACEMA/infection-dating-tool/), and so anyone can deploy their own copy of the tool, or fork the repository and make modifications, as long as the origin of the code is acknowledged and dissemination is also in open source form under the same licensing. The developers welcome contributions to the code. Test characteristics for the more than 60 core HIV diagnostic tests are included in the code base.

As consistent infection dating could be of interest in the study of other infections, a separate version of the system could be deployed to handle other infections, in contexts where multiple diagnostic platforms or algorithms have been used within a single data set intended for a unified analysis. This would naturally involve a completely fresh version of all data in the 'tests' and 'test-properties' tables of the tool's database.

Conclusion

Consistent dating of infection events across subjects has obvious utility when analysing multi-site datasets that contain different underlying screening algorithms. Consistent use of 'diagnostic history' information is also valuable for individual-level interpretation of infection staging at diagnosis.

Even in intensive studies from which 'diagnostic delay' estimates are drawn, it is rarely possible to determine the actual date of infectious exposure. We have adopted a nomenclature based on the earliest date on which an infection would have had 50% probability of being detected, using a viral load assay with a detection threshold of 1 copy per ml, and we refer to this as the Date of Detectable Infection (DDI).

We have presented a simple logic to the interpretation of 'diagnostic testing histories' into 'infection date estimates', either as a point estimate (EDDI) or an interval (EP-DDI – LP-DDI), along with a publicly-accessible online tool that supports wider application of this logic.

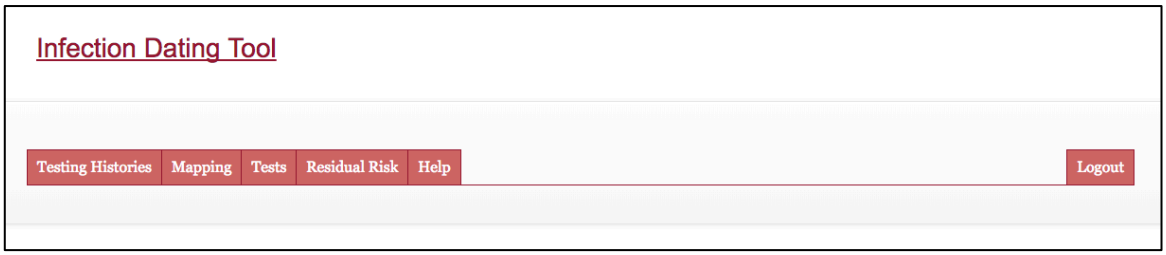
Author contributions

AW: First draft of manuscript, overall project leadership; AW, EG and SF: conceptualisation, data curatorship, code development, writing of the manuscript; RK, MB, GM, CP: conceptualisation; AP, JG, GP, TC: code development.

Appendix: Web interface layout overview

Once logged in, the system presents users with four primary pages, accessible via links spread in horizontal tabs below the header, as shown in Figure A.1. The first three are described in turn below, while the fourth ("Residual Risk") is the subject of a separate publication (Welte *et al*., forthcoming).

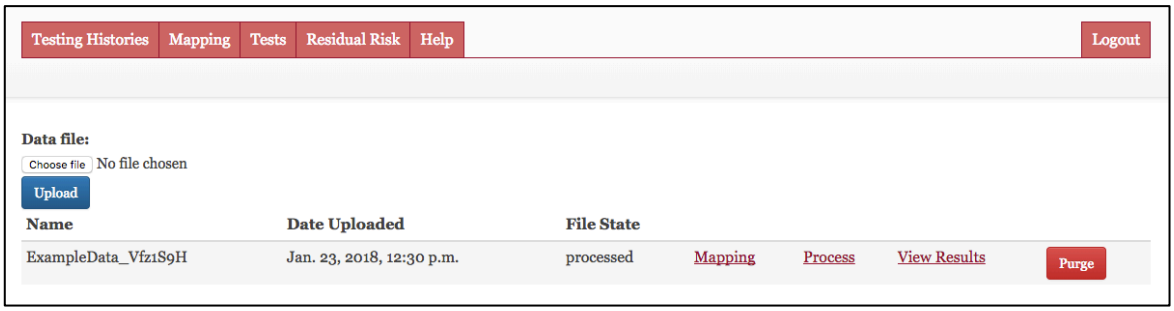
Figure A.1: Navigation



Testing Histories

This tab (Figure A.2) allows users to locate, view and delete previously uploaded 'testing histories', and to upload new ones. It is also where users trigger the action of processing the uploaded testing histories into 'infection dating estimates', which can then be viewed and downloaded.

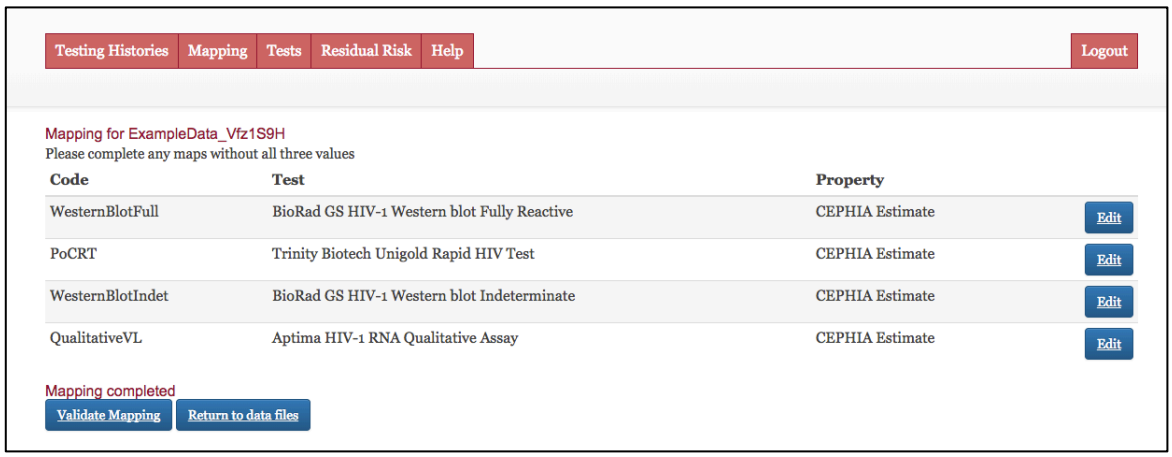
Figure A.2: Testing Histories



Mapping

This tab (Figure A.3) allows users to link strings (alphanumeric codes) in their data files to tests in the online database, hence linking records in uploaded files to the applicable diagnostic delays.

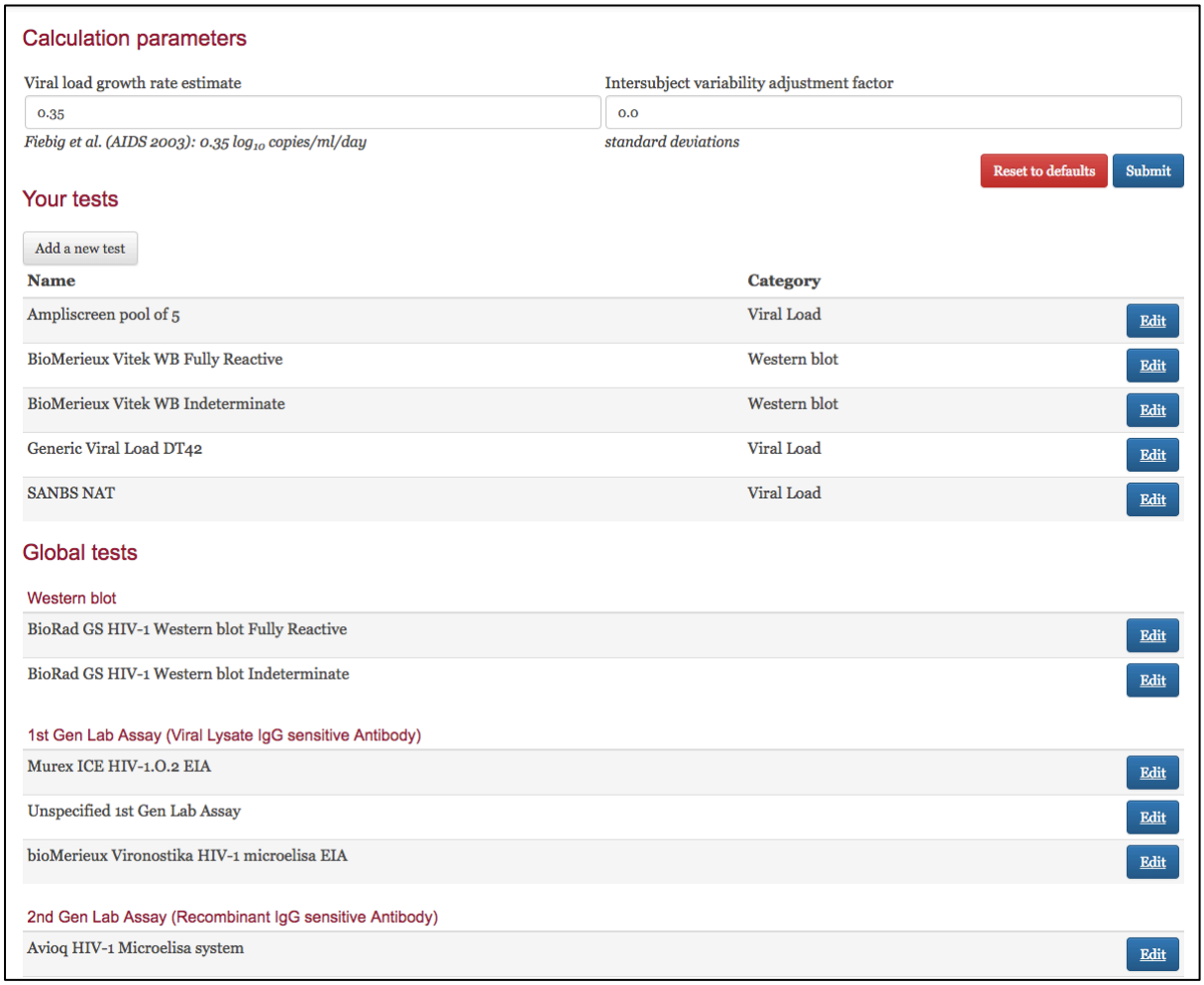
Figure A.3: Mapping



Tests

This tab (Figure A.4) allows users to view the existing database of diagnostic tests, and to add new ones if necessary. Note that each user sees only the shared developer-maintained list of tests, plus his/her own – not those added by other users. This page further allows the user to set a viral load growth rate and to select a multiple of the variability parameter (σ) of the relevant test's diagnostic delay to apply to EP-DDI and LP-DDI estimates as an additional safety margin, as described in the main text.

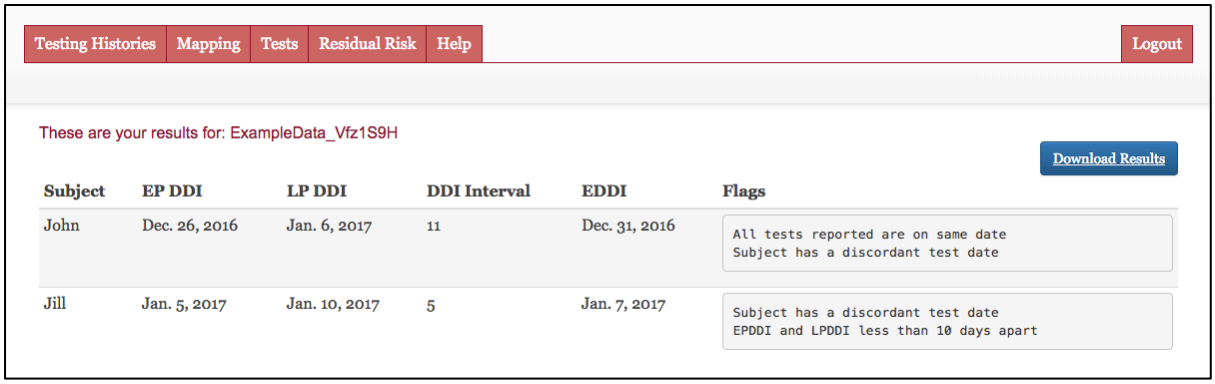
Figure A.4: Tests



Results

Processing can be triggered after test codes have been mapped to specific assays in the database. If test property estimates other than the default are preferred, these can be selected on the mapping screen prior to processing. Each file that has been uploaded on the "Testing Histories" tab has a "Mapping" link, and once mapping has been completed, a "Process" link appears. After processing, results can be viewed and downloaded on a per-file basis (Figure A.5).

Figure A.5: Results



Database Schema

This tool makes use of a relational database, which records information in a set of linked tables, including:

- **subjects:** This table captures each unique study subject, and after infection date estimation has been performed, the subject's EDDI, EP-DDI, LP-DDI and DDI interval size.
- **diagnostic_test_history:** This table records each test performed, by linking to the subjects table and recording a date, a 'test code', and a result. During the estimation procedure, a field containing an 'adjusted date' is populated, which records the candidate EP-DDI (in the case of a negative result) or LP-DDI (in the case of a positive result) after the relevant diagnostic delay has been applied to the actual test date.
- **diagnostic_tests:** This is a lookup table listing all known tests applicable to the current purposes (both system-provided and user-provided).
- **test_property_estimates:** This table records diagnostic delay estimates (system and user-provided). It allows estimates per test, with system default estimates flagged.
- **test_property_mapping:** This table records user-specific mapping of test codes by linking each test code in the diagnostic_test_history table to a test in the diagnostic_tests table, as well as the specific test property estimate 'in use' by that user for the test in question.

A number of subsidiary tables also exist to manage users of the system and allow linking of personal data files, maps, tests, and test property estimates to specific users.

References

1. Fiebig EW, Wright DJ, Rawal BD, Garrett PE, Schumacher RT, Peddada L, et al. Dynamics of HIV viremia and antibody seroconversion in plasma donors: implications for diagnosis and staging of primary HIV infection. Aids. 2003;17(13):1871-9.

2. Lee HY, Giorgi EE, Keele BF, Gaschen B, Athreya GS, Salazar-Gonzalez JF, et al. Modeling sequence evolution in acute HIV-1 infection. Journal of theoretical biology. 2009;261(2):341-60.

3. Ananworanich J, Fletcher JL, Pinyakorn S, van Griensven F, Vandergeeten C, Schuetz A, et al. A novel acute HIV infection staging system based on 4th generation immunoassay. Retrovirology. 2013;10:56.

4. Kassanjee R, Pilcher CD, Keating SM, Facente SN, McKinney E, Price MA, et al. Independent assessment of candidate HIV incidence assays on specimens in the CEPHIA repository. Aids. 2014;28(16):2439-49.

5. Murphy G, Pilcher CD, Keating SM, Kassanjee R, Facente SN, Welte A, et al. Moving towards a reliable HIV incidence test - current status, resources available, future directions and challenges ahead. Epidemiology and infection. 2016:1-17.

6. Kassanjee R, Pilcher CD, Busch MP, Murphy G, Facente SN, Keating SM, et al. Viral load criteria and threshold optimization to improve HIV incidence assay characteristics. Aids. 2016;30(15):2361-71.

7. Kassanjee R. Characterisation and Application of Tests for Recent Infection for HIV Incidence Surveillance. Johannesburg: University of the Witwatersrand; 2014.

8. Kassanjee R, McWalter TA, Barnighausen T, Welte A. A new general biomarker-based incidence estimator. Epidemiology (Cambridge, Mass). 2012;23(5):721-8.

9. Owen SM, Yang C, Spira T, Ou CY, Pau CP, Parekh BS, et al. Alternative algorithms for human immunodeficiency virus infection diagnosis using tests that are licensed in the United States. Journal of clinical microbiology. 2008;46(5):1588-95.

10. Masciotra S, McDougal JS, Feldman J, Sprinkle P, Wesolowski L, Owen SM. Evaluation of an alternative HIV diagnostic algorithm using specimens from seroconversion panels and persons with established HIV infections. Journal of clinical virology : the official publication of the Pan American Society for Clinical Virology. 2011;52 Suppl 1:S17-22.

11. Delaney KP, Hanson DL, Masciotra S, Ethridge SF, Wesolowski L, Owen SM. Time Until Emergence of HIV Test Reactivity Following Infection With HIV-1: Implications for Interpreting Test Results and Retesting After Exposure. Clinical infectious diseases : an official publication of the Infectious Diseases Society of America. 2017;64(1):53-9.