RFU derived LRs for activity level assignments using Bayesian Networks

Peter Gill[a,b], Øyvind Bleka[a], Ane Elida Fonneløp[a]

[a] *Forensic Genetics Research Group, Oslo University Hospital, Oslo, Norway*
[b] *Department of Clinical Medicine, University of Oslo, Oslo, Norway*

Abstract

A comparative study has been carried out, comparing two different methods to estimate activity level likelihood ratios (LR_a) using Bayesian Networks. The first method uses the sub-source likelihood ratio ($log_{10}LR_\phi$) as a 'quality indicator'. However, this has been criticised as introducing potential bias from population differences in allelic proportions. An alternative method has been introduced that is based upon the total RFU of a DNA profile that is adjusted using the mixture proportion (M_x) which is calculated from quantitative probabilistic genotyping software (*EuroForMix*). Bayesian logistic regressions of direct transfer data showed that the two methods were comparable. Differences were attributed to sampling error, and small sample sizes of secondary transfer data. The Bayesian approach facilitates comparative studies by taking account of sampling error; it can easily be extended to compare different methods.

Keywords:
Direct transfer, Secondary Transfer, Mixtures, Likelihood Ratio (LR), Evidence Evaluation, Bayesian Network, ALTRaPht, *EuroForMix*

1. Introduction

In Gill et al. [1], we used sub-source likelihood ratios ($log_{10}LR_\phi$) as a 'quality-indicator' to help assign activity level likelihood ratios (LR_a). Calculations of $log_{10}LR_\phi$s are partly dependent upon population allele frequency databases, with potential bias introduced by rare alleles; secondly, mixture composition may itself have an effect on the $log_{10}LR_\phi$ without affecting the DNA contribution measured by RFU. it was proposed that a measurement

based upon the quantity of DNA from a person of interest (POI) may be preferable. The purpose of this study is to compare such a method with the published $log_{10}LR_{\phi}$ measurement. Probabilities of direct (T) and secondary transfer (S) are modelled by Bayesian logistic regression. An advantage is that sampling uncertainties are taken into account, which is important because sample sizes are small. We show how to compare different methods of measurement.

2. Materials and Methods

Samples were processed and analysed as described in section 2. of [1]; a brief description follows:

1. A total of 33 participants were included in the study in different pair combinations
2. Direct transfer: Participant B contacted the arm of participant A, a minimum of one hour after participant B had washed hands. After the contact, participant A went about normal activities and samples were collected between 0-24 hrs after the initial contact (only one sample collected per experiment). A total of 88 samples were collected.
3. Secondary transfer: A minimum of one hour after washing hands, partipants A and B shook hands for 30sec. and immediately stroked their own upper arm. Samples were taken between 0-24 hrs after contact
4. Powerplex® Fusion 6C multiplex was used to analyse the samples.

2.1. Notation

Details provided in [1] are briefly summarised here:

1. A sample is collected at time h_0. All preceding transfer events are measured relative to the time of sampling.
2. A direct transfer event (T) conditioned under H_p, but disputed by H_d, occurs at time h_1 in hours.
3. Secondary transfer event(s) (S) occur at time h_2 in hours. This may be a range where the number of transfer events (n) are specified.
4. A direct transfer event (T") occurs at time h_3. This is a social interaction such as a handshake and is accepted under H_p and H_d.
5. Activity level LR_a is used generically.
6. Activity level LR_{a_x} is conditioned on $log_{10}LR_{\phi} > x|h_1$ where x is the logistic function decision threshold value.

7. Activity level LR_{a_y} is conditioned on sub-source $RFU_{POI} > y|h_1$, where y is the logistic function decision threshold value.

2.2. Quantitation of the DNA profile from peak height information

From [1], there are two sets of data simulating direct skin to skin contact that may occur in an assault. The first relates to direct transfer experiments, whereas the second data-set relates to the secondary transfer experiments. The data-sets have been expanded (Supplements 4,5) to include details of DNA quantity estimated from the sum of *RFU* values. To carry out this measurement, probabilistic genotyping software *EuroForMix* v.3.3.1 [2] was used to carry out mixture analysis. Mixture proportions (M_x) for the POI are automatically calculated by the software. For a given evidence profile, a script was used to summate all alleles and peaks (RFU_{tot}) above the analytical threshold of $100RFU$, across all (m) STR loci. This was converted into the mean RFU_{tot}/m per locus. Then the contribution to the mean RFU_{POI} was calculated by multiplying RFU_{tot}/m with the POI specific M_x. If the sample is diluted, then multiplication by the dilution factor will compensate the decreased RFU_{POI}, e.g. if one third of the sample has been aliquoted and loaded into the PCR thermocycler, then the dilution factor $d_l = 3$: note that none of the samples in this study were diluted ($d_l = 1$).

$$RFU_{POI} = M_x \times \frac{RFU_{tot}}{m} \times d_l \tag{1}$$

2.3. Bayesian logistic regression

All programming was carried out in R v. 4.0.3 [3]. To analyse the direct transfer results, Bayesian logistic regression, using the R *stan_glm* function [4] from package (*rstanarm*), utilising MCMC to generate 4000 pairs of coefficients per logistic regression to find $Pr(RFU_{POI} > y|h_1)$ where y is the threshold value tested and h_1 is the time-interval in hours between the transfer and collection of samples .

2.4. Exponential distribution fits to secondary transfer data

The combined data between time zero and one hours after deposition were modelled using an exponential distribution to determine $Pr(S|h_2 \in 0, 1)$. The exponential distribution was determined to be the best fit after testing different options with the R package (*fitdistrplus*) [5] (supplement S3; Fig. S2).

2.5. Likelihood ratio calculations

Activity level LR_{a_y} was calculated using a modified Activity Level, Transfer, Results and Persistence program [1] (ALTRaPht); source files available at `https://sites.google.com/view/altrap/home`, along with the datasets and user manual. The program can be launched directly from the 'shinyapps.io' server site `https://activitylevel.shinyapps.io/shinyaltrapht/` without any need to load R.

3. Results

3.1. Linear regression

Linear regression of the direct transfer data (where $log_{10}LR_\phi > 0$ vs RFU_{POI}) was carried out (Fig. 1); Intercept=21.7; Slope=24.6. $R^2 = 0.8$. Higher values are typically associated with short time intervals between DNA transfer and sampling.

3.2. Logistic regression

Linear regression does not take into account the time since transfer (h), and was insufficient to fully characterise the relationship of RFU and $log_{10}LR_\phi$ relative to h. In the next step, logistic regressions of $RFU > y|h_1$ were fitted to regressions of $log_{10}LR_\phi > x|h_1$. This was carried out by optimisation as follows:

Step 1: A table of logistic regression probabilities for $x = 1, .., 12$ was prepared (Supplement 6).

Step 2: To find the equivalent of $RFU_{POI} > y|h_1$ for a given x threshold (Table 1), optimisation was carried out in order to determine the values of y that gave the best fit to $log_{10}LR_\phi > x|h_1$ logistic regressions. This was carried out by iterating y for a specific value of x in order to minimise the mean square error (MSE) i.e. the mean of the sum of squared residuals across all time points in hours (0,1,...,24). The resultant logistic regressions were all good fits (Fig 2; supplement S2; Table S1). The "Optimizer.R" program used to carry out the optimisation is available at `https://github.com/peterdgill/altrapHt.git`.

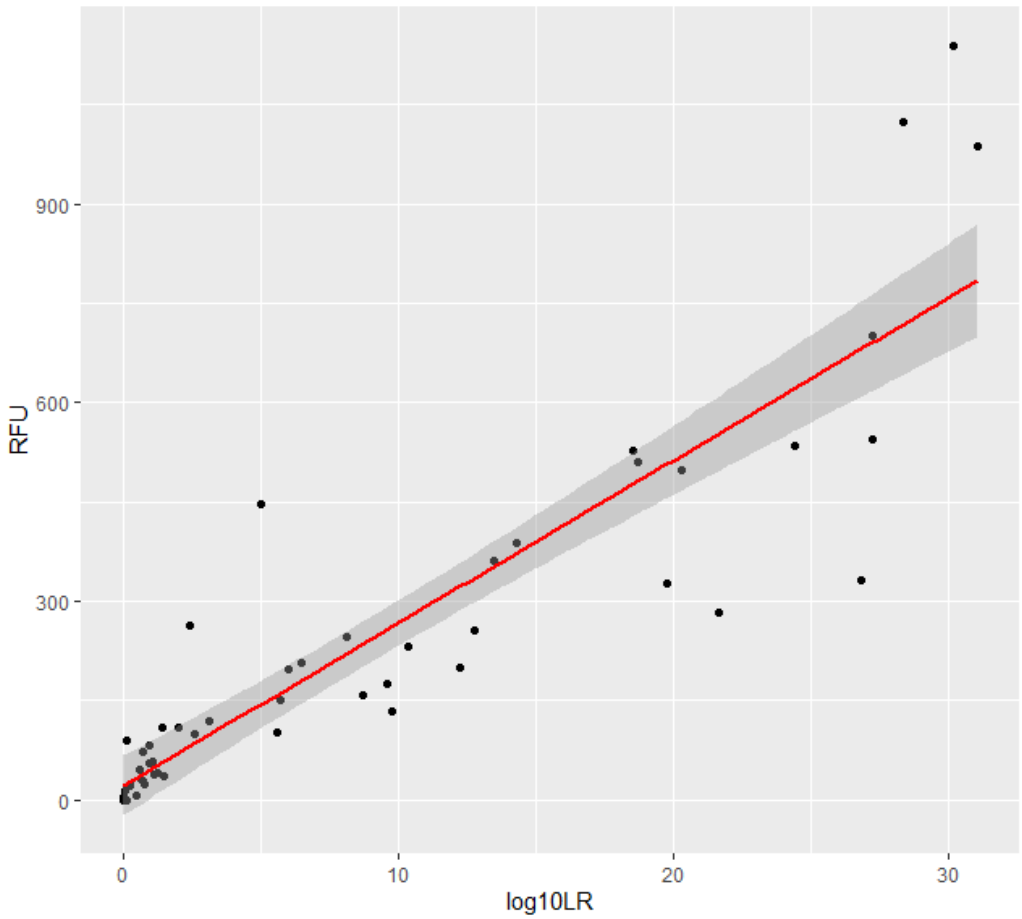

Figure 1: Linear Regression analysis of direct transfer data $log_{10}LR_{\phi}$ vs. M_x adjusted mean peak height of the POI (RFU_{POI})

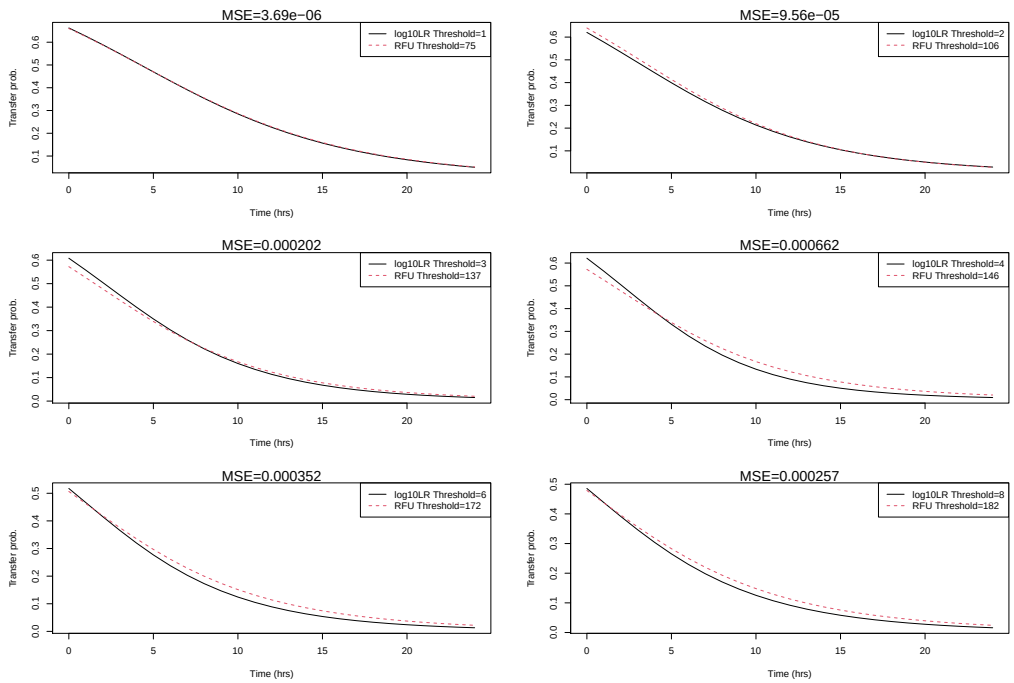

Figure 2: Logistic Regression analysis of direct transfer data, Time vs. Transfer probability, where each panel compares $log_{10}LR_{\phi} > x|h_1$ with $RFU_{POI} > y|h_1$ equivalents listed in Table 1

$log_{10}LR_\phi > x$	$RFU > y$ equivalent	MSE
1	75	3.0E-06
2	106	1.0E-05
3	137	2.0E-04
4	146	7.0E-04
6	172	3.0E-04
8	182	3.0E-04
10	237	5.0E-04
12	266	1.0E-04

Table 1: $RFU_{POI} > y$ equivalents to $log_{10}LR_\phi > x$ logistic regression thresholds, with mean square error (MSE) of the residuals

Step 3: The Activity Level, Transfer Recovery and Persistence (ALTRaP) program [1] was modified to accept RFU_{POI} values instead of $log_{10}LR_\phi$ values.

Some further data exploration using optimisation is described in Supplement S1.

4. Discussion

4.1. Reworked example

In Gill et al. [1], the ALTRaP program was used to calculate activity level LR_{a_x} based upon sub-source $log_{10}LR_\phi$ values. To illustrate the method a number of examples were described. Details of the first example are briefly summarised as follows: Sampling was at time $h = 0$, following an assault at time $h_1 = 3$ and social contact at $h_2 = 6, ..., 10$ (one contact by secondary transfer each hour, i.e. $n = 1$); direct innocent contact e.g. handshake at time $h_3 = 10$. Activity level LR_{a_x} vs. LR_{a_y}, were calculated and compared (Table 2).

Activity level LR_as are comparable for logistic thresholds $x = 1, 2, 3$ and y equivalents, but for higher values of x, the LR_{a_y} results plateau. Logistic regression coefficients for direct transfer are similar when $x > 8$ [1], because there are few observations, especially when $h_1 > 9$ hrs.

$Pr(E\|Hp)$ proposition	Type of transfer		T and S		(T, S and T")	
	$log_{10}LR_\phi > x$	$RFU > y$	LR_{a_x}	LR_{a_y}	LR_{a_x}	LR_{a_y}
POI only	1	75	6	8	3	4
	2	106	10	12	4	5
	3	137	18	16	5	5
	4	146	33	19	6	5
	6	172	43	22	5	5
	8	182	72	24	5	4
POI and unknown(s)	1	75	1	2	1	1
	2	106	3	3	1	1
	3	137	5	5	1	1
	4	146	10	6	2	2
	6	172	16	8	2	2
	8	182	29	9	2	2

Table 2: Comparison of activity level LR_{a_x} and LR_{a_y} using logistic regression thresholds $log_{10}LR_\phi > x$ and the equivalent $RFU_{POI} > y$. Direct transfer (T) $h_1 = 3$; secondary transfer (S) $h_2 = 6, .., 10$, with one contact per hour ($n = 1$); direct (innocent) transfer (T") $h_3 = 10$. For full details see Gill et al [1] Table 3

4.2. Investigation into effect of small sample size

An investigation was carried out to determine the reasons for the divergences between LR_{a_x} and LR_{a_y} values. LR_a is calculated from Bayesian logistic regression analysis of direct and secondary transfer data. Because of the very low success rate of secondary transfer, especially for time intervals > 1hr, only data immediately following the secondary transfer opportunity was analysed ($h_2 = 0, 1$) as described in section 8 of Gill et al. [1], except that data were fitted to an exponential distribution rather than Pareto. Preferred distributions, and associated parameters, were determined by diagnostic tests using the *fitdistrplus* R-package (Supplement S3; Fig. S2).

From Bayesian logistic regressions, quantiles were tabulated in rows (section 14.5 of [1]), corresponding to quantiles derived from 1000x bootstraps of the secondary transfer probabilities $Pr(S)$ (see section 14.6 of [1]). Briefly: data collected from zero - one hour after the secondary transfer experiment, were subject to bootstrapping [6] in order to generate 1000 new randomised sets of samples. For each data-set, new exponential distribution parameters were estimated, and $Pr(S|y)$ was determined as described in [1] (section 8 and fig 6). Quantiles that describe the range of $Pr(S|y)$ were generated from

the 1000x bootstrapped samples. The second step was to substitute a value of $Pr(S|y)$ in the Bayesian logistic regression calculation, where 4000 pairs of data-dependant simulated coefficients were generated in order to calculate likelihood ratios. These data are also described by quantiles. Following these two steps, a plausible range of likelihood ratios can be prepared. In particular, we are interested in the range of LR_{a_y} defined by 0.25-0.75 quantiles (Table 3) which can be described as encompassing the plausible range of activity level likelihood ratios.

Table 3 shows an analysis for the example described in Table 2: the grey cells are median values of LR_{a_y}; the orange cells show the nearest values associated with the LR_{a_x} for a given logistic regression threshold (x) that lies within the 0.25-0.75 quantiles. Greater divergence is observed when $x = 8$; $y = 182$, but there are few values observed in the data sets. In conclusion, the differences in LR_{a_x} vs. LR_{a_y} can be explained by sampling error since their respective values fall within the 0.25-0.75 quantiles of direct transfer logistic regressions and secondary transfer bootstrapped samples.

4.3. Reporting

Here we have described two alternative methods to assign LR_a from the same data. Differences may be attributed to sampling.

There are a number of ways that we could proceed: the simplest option would be to adopt a single method, i.e. the most informative method chosen after further comparative studies. Using both methods concurrently is possible, but this raises the question of how to report? As an analogy with probabilistic genotyping, it is useful to refer to the debate arising from [7, 8] and subsequent discussions by Taylor et al. [9], where the latter suggest reporting an average LR where two or more (comparable) methods are employed. When sensitivity analysis is carried out, reporting the median value has a good consensus - see discussion in [1], section 14.9.

As explained by Taroni et al. [10], the issue is not uncertainty of the likelihood ratio, rather it is a question of precision and communication. The use of 0.25 -0.75 quantiles can be used as a sensitivity analysis to provide a plausible range of LR_a values. The authors suggest that one way to simplify communication, would be to avoid the use of numbers completely. However, this idea runs counter to the ISFG DNA commission [11] which recommends that the numeric likelihood ratio as compulsory, and the verbal analogue as optional. The precision of the results from sensitivity analysis suggests that

		logistic regression quantiles		
logistic regression threshold	secondary transfer quantiles	**0.25**	**0.5**	**0.75**
$log_{10}LR_{\phi} > 4$; $146RFU$	**0.25**	22	**31**	**43**
	0.50	14	**19**	27
	0.75	10	14	19
$log_{10}LR_{\phi} > 6$; $172RFU$	**0.25**	29	**41**	**59**
	0.50	16	**23**	32
	0.75	11	16	22
$log_{10}LR_{\phi} > 8$; $182RFU$	**0.10**	59	**82**	119
	0.25	30	42	60
	0.50	17	**23**	33
	0.75	11	16	22

Table 3: Secondary transfer data were bootstrapped to generate 1000 new randomised sets of samples. Exponential distributions were estimated for each to determine $Pr(S|x)$ and the range was expressed by quantiles in the rows of the table. Respective $Pr(S|x, y)$ values were plugged into Bayesian logistic regressions that generated the quantiles denoted by the columns. Figures highlighted in grey are the median LR_{a_y} results. The figures highlighted in orange indicate the LR_{a_x} equivalent values

the reported value of LR_a should be rounded down to multiples of ten, unless it is less than ten.

In supplement S4 further examples are explored, taken directly from [1]. The trend for lower LR_{a_y} compared to LR_{a_x} values is repeated. If this is mainly attributable to sampling error of secondary transfer data-sets, then larger sample sizes should improve the comparability of the results and reduce the observed variance, thereby increasing precision.

4.4. Conclusion

This is part of a study to lay the ground-work for a much larger project which will begin in 2022, entitled: *"Establishment of a Trace DNA Transfer Rate Repository and Bayes Nets to Calculate LRs (ReAct)"*. Funded by the European Network of Forensic Science Institutes (ENFSI) monopoly project, there are more than 30 participating laboratories. This will provide an opportunity to generate new sets of data that will help to optimise the methodology and experimental design, in addition to facilitating much needed tests of reproducibility. It is anticipated that the "Optimizer.R" program may be useful to compare logistic regressions from different laboratories to determine equivalence factors as described in section 3.2. Taking the mean RFU measurement across loci facilitates comparisons across different multiplexes. To measure M_x, a laboratory will need to use qualitative (continuous) probabilistic genotyping software. Comparative studies are needed to address complex issues related to standardisation: where the aim is to show whether different laboratories following similar methods, report similar LR_as for given conditions.

References

[1] P. Gill, Ø. Bleka, A. Roseth, A. E. Fonneløp, An LR framework incorporating sensitivity analysis to model multiple direct and secondary transfer events on skin surface, Forensic Science International: Genetics 53 (2021) 102509.

[2] Ø. Bleka, G. Storvik, P. Gill, *EuroForMix*: an open source software based on a continuous model to evaluate STR DNA profiles from a mixture of contributors with artefacts, Forensic Science International: Genetics 21 (2016) 35–44.

[3] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria (2013). URL http://www.R-project.org/

[4] C. Muth, Z. Oravecz, J. Gabry, User-friendly bayesian regression modeling: A tutorial with rstanarm and shinystan, Quantitative Methods for Psychology 14 (2) (2018) 99–119.

[5] M. L. Delignette-Muller, C. Dutang, et al., fitdistrplus: An R package for fitting distributions, Journal of statistical software 64 (4) (2015) 1–34.

[6] B. Efron, R. J. Tibshirani, An introduction to the bootstrap, CRC press, 1994.

[7] E. Alladio, M. Omedei, S. Cisana, G. D'Amico, D. Caneparo, M. Vincenti, P. Garofano, DNA mixtures interpretation–a proof-of-concept multi-software comparison highlighting different probabilistic methods' performances on challenging samples, Forensic Science International: Genetics 37 (2018) 143–150.

[8] P. Garofano, D. Caneparo, G. D'Amico, M. Vincenti, E. Alladio, An alternative application of the consensus method to DNA typing interpretation for low template-DNA mixtures, Forensic Science International: Genetics Supplement Series 5 (2015) e422–e424.

[9] D. A. Taylor, J. S. Buckleton, J.-A. Bright, Comment on "DNA mixtures interpretation–a proof-of-concept multi-software comparison highlighting different probabilistic methods' performances on challenging samples" by Alladio et al., Forensic Science International: Genetics 40 (2019) e248–e251.

[10] F. Taroni, S. Bozza, A. Biedermann, C. Aitken, Dismissal of the illusion of uncertainty in the assessment of a likelihood ratio, Law, Probability and Risk 15 (1) (2016) 1–16.

[11] P. Gill, T. Hicks, J. M. Butler, E. Connolly, L. Gusmão, B. Kokshoorn, N. Morling, R. A. van Oorschot, W. Parson, M. Prinz, et al., DNA commission of the international society for forensic genetics: Assessing the value of forensic biological evidence-guidelines highlighting the importance of propositions: Part I: evaluation of DNA profiling comparisons

given (sub-) source propositions, Forensic Science International: Genetics 36 (2018) 189–202.

Supplementary material

S1. Relationship of $\boldsymbol{Pr(log_{10}LR_{\phi} > x|h)}$ vs. $\boldsymbol{Pr(RFU_{POI} > y|h)}$ explored by optimisation

To further explore the relationship of $Pr(log_{10}LR_{\phi} > x|h)$ vs. $Pr(RFU_{POI} > y|h)$, optimisation was carried out to find the equivalent of $RFU_{POI} > y|h_1$ for a given x logistic regression threshold, exactly as described in step 2 of section 3.2, except that the entire range of RFU values was explored. Next, complementary calculations were made to estimate the equivalent of $log_{10}LR_{\phi}|h_1$ for a given y threshold. The results (Fig. S1 showed a linear regression for the range $log_{10}LR_{\phi}$=1-20 ($R^2 = 0.97$) and the equivalent RFU=50-400. There are demonstrable upper limits for both sets of calculations: the first limit is at $450RFU$ and the second is at $log_{10}LR_{\phi} = 19$. Note that the majority of data for direct transfer and all secondary transfer data gave $log_{10}LR_{\phi} < 10$ and $RFU < 250$. Some high LRs or RFUs may be observed for short time intervals between sample deposition and collection, but for casework, short time intervals are usually unrealistic. consequently, we focus the discussion based upon the range of results that best represent typical casework expectations. This analysis did not determine which of the two measures was preferable to use.

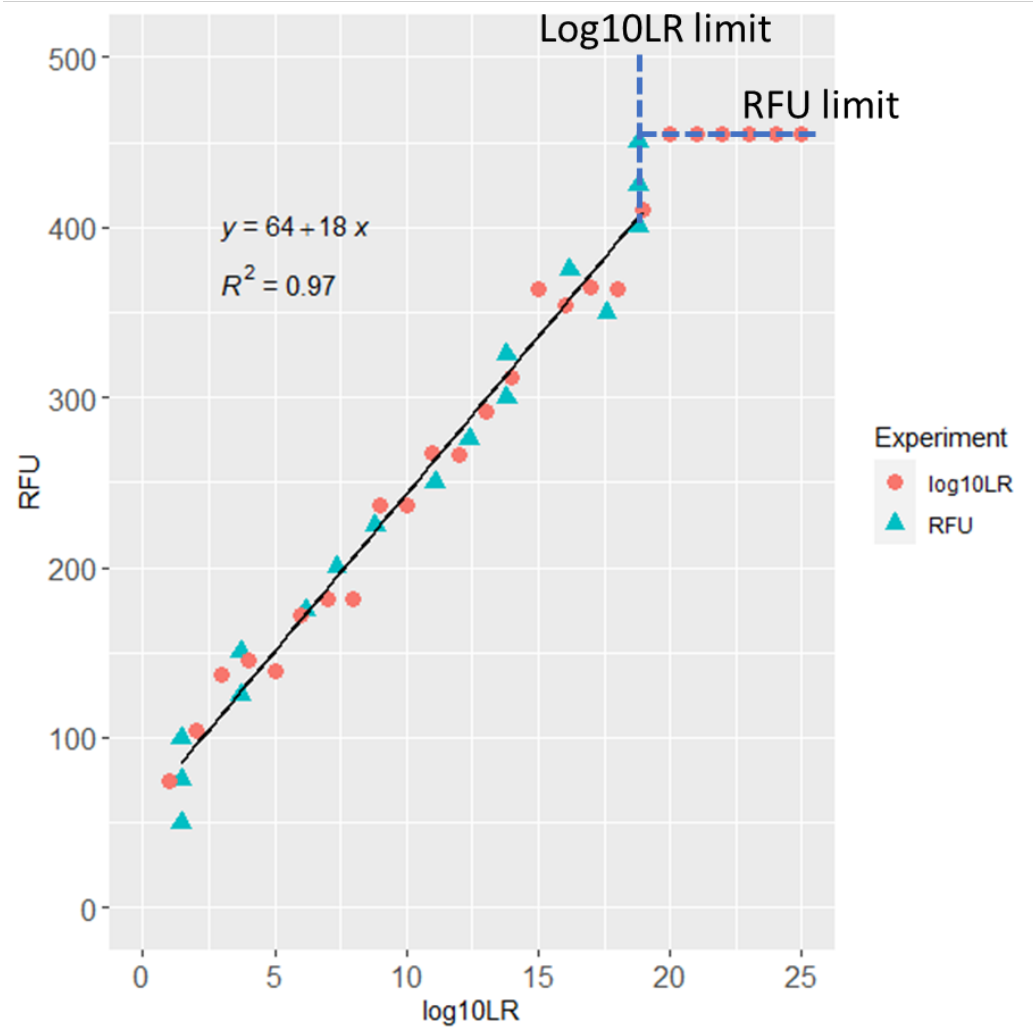

Figure S1: $RFU_{POI} > y$ equivalents to given $log_{10}LR_{\phi} > x$ logistic regression thresholds and complementary $log_{10}LR_{\phi} > x$ equivalents to given $RFU_{POI} > y$ thresholds. The conditioned variable is shown in the key

S2. Summary statistics for logistic regression models

S2.1. Summary statistics

1. n_eff is an estimate of the effective number of independent draws from the posterior distribution of the estimand of interest

	Rhat	n.eff	mean	sd	se_mean	2.50%	25%	50%		75%	97.50%
$y = 75; x = 1$											
(Intercept)	1.0002	2,404	0.6853	0.3694	0.0075	-0.0158	0.4356	0.6774	**(0.5690)**	0.9256	1.4247
Time	1.0014	1,963	-0.1614	0.0452	0.001	-0.2584	-0.1904	-0.1592	**(-0.1490)**	-0.1297	-0.0808
$y = 106; x = 2$											
(Intercept)	1.0014	2,870	0.5938	0.361	0.0067	-0.1066	0.3448	0.5903	**(0.5084)**	0.8286	1.3215
Time	1.0032	1,751	-0.1895	0.0505	0.0012	-0.2954	-0.2226	-0.1864	**(-0.1823)**	-0.1542	-0.0993
$y = 137; x = 3$											
(Intercept)	1.001	2,586	0.3016	0.3695	0.0073	-0.3903	0.0449	0.2929	**(0.4430)**	0.5499	1.0595
Time	1.0017	1,558	-0.1961	0.0577	0.0015	-0.3222	-0.2324	-0.1918	**(-0.2120)**	-0.1569	-0.0941
$y = 146; x = 4$											
(Intercept)	1.001	2,586	0.3016	0.3695	0.0073	-0.3903	0.0449	0.2929	**(0.5075)**	0.5499	1.0595
Time	1.0017	1,558	-0.1961	0.0577	0.0015	-0.3222	-0.2324	-0.1918	**(-0.2431)**	-0.1569	-0.0941
$y = 139; x = 5$											
(Intercept)	1.001	2,586	0.3016	0.3695	0.0073	-0.3903	0.0449	0.2929	**(0.0783)**	0.5499	1.0595
Time	1.0017	1,558	-0.1961	0.0577	0.0015	-0.3222	-0.2324	-0.1918	**(-0.2084)**	-0.1569	-0.0941
$y = 172; x = 6$											
(Intercept)	1.0004	2,669	0.0274	0.3591	0.007	-0.6659	-0.221	0.0242	**(-0.0517)**	0.2668	0.7492
Time	1.0027	1,371	-0.1809	0.056	0.0015	-0.2985	-0.2175	-0.177	**(-0.1951)**	-0.1409	-0.0827
$y = 182; x = 8$											
(Intercept)	1.0004	2,582	-0.0868	0.3622	0.0071	-0.7883	-0.3364	-0.0781	**(-0.0517)**	0.1535	0.6303
Time	1.0005	1,413	-0.1719	0.0561	0.0015	-0.2914	-0.2083	-0.1692	**(-0.1951)**	-0.1319	-0.0727

Table S1: Summary statistics of Direct transfer Bayesian logistic regression statistics generated from 4000 simulations per test. For each decision threshold $RFU_{POI} > y$, there are two coefficients (Intercept and Time) that are used to generate the curves in Fig 2. The $log_{10}LR_{\phi} > x$ threshold equivalents are also provided in parentheses. Summary statistics are discussed in section S2.1. Percentiles for each coefficient are also listed. The 50 percentiles are compared with that for $log_{10}LR_{\phi} > x$ equivalents in parentheses.

2. Rhat: One way to monitor whether a chain has converged to the equilibrium distribution is to compare its behavior to other randomly initialized chains. This is the motivation for the Gelman and Rubin potential scale reduction statistic Rhat. The Rhat statistic measures the ratio of the average variance of samples within each chain to the variance of the pooled samples across chains; if all chains are at equilibrium, these will be the same and Rhat will be one. If the chains have not converged to a common distribution, the Rhat statistic will be greater than one.
3. se_mean: The standard error of the mean of the posterior draws (not to be confused with the standard deviation of the posterior draws) is the uncertainty associated with the Monte Carlo approximation. This quantity approaches 0 as the sample size goes to infinity.
4. sd: The posterior standard deviation of the posterior draws
5. The 2.5, 50 and 97.5 percentiles are shown

S3. Exponential distribution calculation

S3.1. Exponential distribution calculation

For secondary transfer, zero and one hour after deposition, the proportion of data (k) where $RFU = 0$, were excluded from the initial distribution fitting (fig S2); $k = 0.5$. This greatly improved the fit without affecting probabilistic calculations of $Pr(RFU > y)$, since the logistic regression decision threshold y is always greater than zero.

We modelled $Pr(RFU = y)$ as a mixture model ($y = 0$ vs. $y > 0$), where $Pr(RFU = 0) = k$, and an exponential distribution weighted with $(1 - k)$ for $y > 0$, was used to define the probability of secondary transfer. Hence the cumulative expression (when $y > 0$) is given as:

$$Pr(RFU <= y) = k + (1 - k)(1 – e^{-\lambda y})$$

The derivation of $Pr(RFU > y)$ (eq S1) is shown below:

$$\begin{aligned} Pr(RFU > y) &= 1 - (k + (1 - k) \overbrace{(1 - e^{-\lambda y})}^{\text{exp. distribution}}) \\ &= 1 - (k + 1 - e^{-\lambda y} - k + ke^{-\lambda y}) \\ &= e^{-\lambda y} - ke^{-\lambda y} \end{aligned}$$

$$Pr(RFU > y) = (1 - k) \left(e^{-\lambda y}\right) \tag{S1}$$

The probability of direct transfer is defined by the logistic regression in eq S2:

$$Pr(T_h) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 h)}} \tag{S2}$$

where β_0 and β_1 are the Intercept and Time coefficients respectively.

We seek to find (from eq: S3):

$$Pr(T_{h=f_y}|y) = Pr(S_{h=0}|y) \tag{S3}$$

Where f_y is the offset value, which we need to determine from the following expression:

$$\frac{1}{1 + e^{-(\beta_0 + \beta_1 f_y)}} = (1 - k) \left(e^{-\lambda y}\right) \tag{S4}$$

By rearrangement, to calculate f_y:

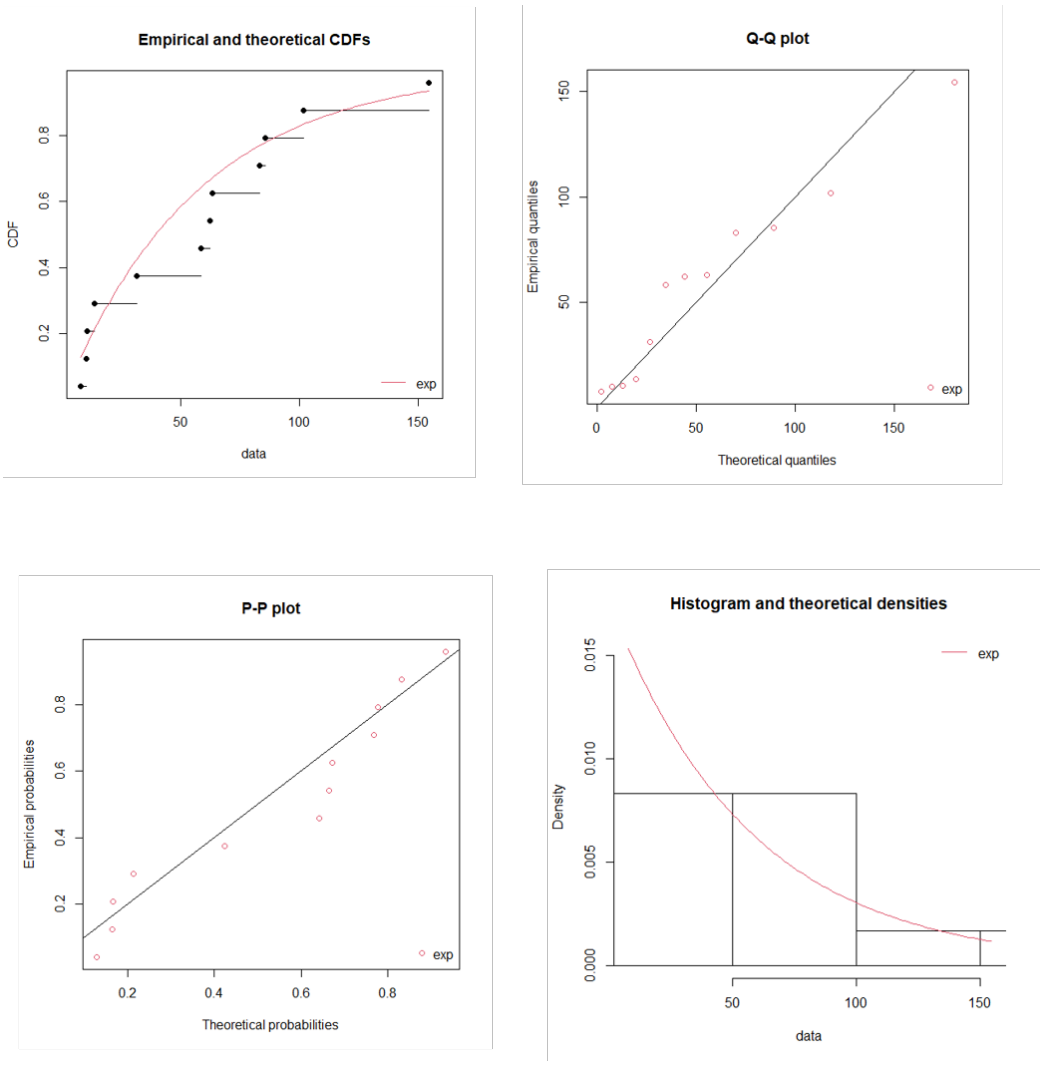

Figure S2: Exponential distribution fit to secondary transfer data (zero and one hour after deposition) showing the CDF, and diagnostic Q-Q and P-P plots. Zero RFU data were excluded from the analysis. The R-package *fitdistrplus* was used to generate the plots.

$$f_y = -\frac{1}{\beta_1}\left(log\left(\frac{1}{1-k} \times e^{\lambda y} - 1\right) - \beta_0\right) \tag{S5}$$

This expression is encoded into ALTRaP and is used to determine the offset value for each of the 4000 simulations per value of y.

Note that for the $Pr(S)$ sensitivity tab in ALTRaP, the value is determined by optimisation.

S4. Reworked examples from [1]

The following example replicates Table 4 from [1]. Varying times since direct h_1 and secondary h_2 transfer are shown in columns 1 and 2 of Table S2 along with the respective LR_{a_x} and LR_{a_y} values. For each calculation the threshold values are the same: $x = 3$; $y = 137$ (from Table 1). Sensitivity analysis of a few selected examples are shown in Table S3. LR_{a_x} values are within range of 0.25 LR_{a_y} secondary transfer quantiles.

In Table S3, sensitivity analysis of some examples in Table S2 has been undertaken. 0.25-0.75 quantiles are shown undertaken for LR_{a_y} data; medians are shown in grey and LR_{a_x} equivalents are shown in orange. Note that all results are within the 0.25 secondary transfer quantiles.

In Table S4 multiple secondary transfer events (h_2) are modelled ($n = 1$). Respective LR_{a_x} and LR_{a_y} values are compared. For each calculation the threshold (x) values are shown (see Table 1 for y_{RFU} equivalents. Sensitivity analysis of a few selected examples are shown in Table S5. LR_{a_x} values are within range of 0.25 secondary transfer LR_{a_y} quantile and 0.5-0.75 logistic regression quantiles

		POI only		POI + unknown(s)	
Time since direct transfer (h_1)	Time since secondary transfer (h_2)	LR_{a_x}	LR_{a_y}	LR_{a_x}	LR_{a_y}
6	8	41	39	18	18
6	10	62	56	27	26
6	14	168	118	70	54
6	18	329	251	149	116
8	10	41	39	22	21
8	12	62	56	34	31
8	16	142	118	78	65
8	20	329	251	183	139
12	14	41	39	30	28
12	16	61	56	45	40
12	20	142	118	104	85
12	24	329	251	244	182

Table S2: Analysis of the second example described in Gill et al. [1] (Table 4), comparing LR_{a_x} and LR_{a_y} $x = 3$; $y = 137RFU$; number of secondary transfer contacts per hour is n =1. Time of direct transfer e.g. assault is h_1 and time of secondary transfer e.g. from social contact is h_2

			logistic regression quantiles		
	h_1; h_2	secondary transfer quantiles	**0.25**	**0.5**	**0.75**
POI only	$h_1 = 6$; $h_2 = 18$	**0.25**	226	**412**	765
		0.50	142	**258**	479
		0.75	105	189	362
POI + unknown(s)	$h_1 = 6$; $h_2 = 18$	**0.25**	102	**186**	351
		0.50	64	**116**	219
		0.75	48	87	168
POI + unknown(s)	$h_1 = 6$; $h_2 = 24$	**0.25**	150	**293**	578
		0.50	94	**183**	361
		0.75	66	129	253

Table S3: Three examples taken from Table S2 where sensitivity analysis has been carried out. The grey cells are the median values of LR_{a_y} and the orange cells correspond to nearest LR_{a_x} values.

			POI only		POI and unknown	
h_1	h_2	$log_{10}LR_\phi > x$	LR_{a_x}	LR_{a_y}	LR_{a_x}	LR_{a_y}
		1	3	4	1	1
		2	5	6	1	1
0	0,12	3	7	7	1	2
		4	12	8	2	2
		6	15	9	4	3
		8	25	9	7	3
		1	4	5	1	1
		2	7	8	2	2
3	6,18	3	12	10	4	3
		4	22	12	7	4
		6	27	13	10	5
		8	44	14	18	5
		1	5	7	2	2
		2	10	12	4	5
6	12,24	3	20	17	9	8
		4	45	19	21	9
		6	50	21	26	11
		8	78	22	42	11
		1	22	39	7	12
		2	81	105	32	40
6	24,36	3	246	156	112	71
		4	807	191	388	89
		6	593	192	312	98
		8	799	161	432	84

Table S4: Analysis of the example described in [1], Table 5, where multiple secondary transfer events, over a 12 hour time interval given by h_2,with $n = 1$ transfer event per hour. LR_{a_x} is compared with the equivalent LR_{a_y}

			logistic regression quantiles					
logistic threshold	h_1;h_2	Secondary transfer quantiles	POI only			POI and unknown		
			0.25	**0.5**	**0.75**	**0.25**	**0.5**	**0.75**
		0.25	109	297	**807**	56	15	**422**
x=8; y=182RFU	$h_1 = 6$;$h_2 = 24, .., 36$	**0.50**	60	**162**	440	31	**84**	231
		0.75	39	105	285	20	54	149
		0.25	25	**41**	**71**	12	**21**	36
x=6; y=172RFU	$h_1 = 6$;$h_2 = 12, .., 24$	**0.50**	14	**22**	37	7	**11**	19
		0.75	9	15	25	5	7	13

Table S5: Two examples taken from Table S4 where sensitivity analysis has been carried out. The grey cells are the median values of LR_{a_y} and the orange cells correspond to nearest LR_{a_x} values.