Follow the trace

Evaluating physical and digital forensic findings given activity level propositions using Bayesian networks

Vink, M.

Publication date
2025

Chapter 6

AN LR METHOD FOR THE INTERPRETATION OF IPHONE HEALTH APP DATA

Based on[1] Vink, M., Sjerps, M. J., Boztas A., & Van Zandwijk J. P. 2022. Likelihood ratio method for the interpretation of iPhone Health App data in digital forensics. *Forensic Science International: Digital Investigation* **41**, 301389[2].

[1]This paper has been published with subscription access. Minor modifications have been made to the publication: see Page 289 for a clarification of the modifications.

[2]The authors' first names were incorrectly formatted in the official citation. The official citation is: Vink, M. M., Sjerps, M. J. M., Boztas, A. A., & Van Zandwijk, J. P. J. P. 2022. Likelihood ratio method for the interpretation of iPhone Health App data in digital forensics, *Forensic Science International: Digital Investigation* 41, 301389.

6.1 Introduction

Forensic science largely concerns drawing inferences from observations and/or measurements in a forensic setting. Until the 21st century, almost all of these inferences related to the source of evidentiary material. This source level evaluation is used to conceivably relate trace material to a source and to link individuals, objects and/or locations accordingly (Cook *et al.*, 1998). Forensic evidence may also be evaluated at activity level. Activity level evaluations become more common as these evaluations address the question: "What activity has led to the deposition of the trace?" A paradigm shift within forensic science is therefore foreseeable. Yet, activity level evaluation is still in its early stages in most forensic disciplines - digital forensic science being the latest example.

Digital forensic science focuses on the recovery and analysis of digital traces left by individuals on digital devices. Individuals have become more attached to and reliant on digital devices such as laptops and smartphones. Consequently, digital traces are strongly connected to an individual's activity and/or location, resulting in an increased overlap between digital traces and the physical world. Because digital traces often contain detailed timestamps, it follows that they are increasingly valuable in criminal cases and likewise in activity level evaluations (Henseler & De Poot, 2020; Van Zandwijk & Boztas, 2019, 2021). A few papers have been published on the forensic evaluation of event data in the past few years, all illustrating the possibility and importance of balanced evidence evaluation in digital forensic science (Bosma *et al.*, 2020; Casey *et al.*, 2020; Galbraith *et al.*, 2020a,b).

Most smartphones and smartwatches are equipped with sensors and applications that monitor health related activities such as heart rate and step count in time, introducing numerous opportunities for digital forensic science. This study focuses on step count data and distance registrations from the iPhone Health app. Several studies (Åkerberg *et al.*, 2016; Duncan *et al.*, 2018; Höchsmann *et al.*, 2018, 2020; Van Zandwijk & Boztas, 2019) investigate the accuracy and precision of step count and distance data (hereafter: step count data) gathered by different applications and conclude that the applications are suitable tools for step counting and distance registration. As such, step count data may add valuable information to criminal investigations as they provide a link between time and physical activity of the user. In fact, Van Zandwijk & Boztas (2019) recognize the forensic application of step count data and argue that step count data from the iPhone Health app can be used to make probability statements on an individual's alleged walking route.

In addition, the Dutch court acknowledges that step count data can be of help in criminal investigations, causing step count data to be utilized increasingly in the last few years. Step count data has been used to either confirm an individual's movement[3] or to analyze an individual's alleged activity[4]. In the latter, the step count is translated to a certain distance which is then used to state whether the data is consistent with the alleged activity. Statements on consistency have been categorical and accordingly none of these cases used step count data within a likelihood ratio (LR) framework. Still, the LR framework is the preferred method of inference: it presents the degree of support for one hypothesis versus another and provides a logically correct method to assist the expert in assessing their uncertainty (Aitken *et al.*, 2020; Evett, 1998, 2015; Robertson *et al.*, 2016; Taroni *et al.*, 2002). Moreover, the LR framework clarifies the expert's evaluative role and the jurors' decision-making role.

Even though the accuracy and precision of step count data has been studied extensively, research on how to use this data within an LR framework has, to the best of our knowledge, not been published yet. This study contributes to digital forensic science by presenting a numerical LR method for experimentally

[3]Rb. Den Haag 13 November 2019, ECLI:NL:RBDHA:2019:11950.
Rb. Gelderland 21 January 2020, ECLI:NL:RBGEL:2020:281.
Rb. Noord-Nederland 23 November 2020, ECLI:NL:RBNNE:2020:3975.
Hof Arnhem-Leeuwarden 1 December 2020, ECLI:NL:GHARL:2020:9865.
[4]Rb. Den Haag 5 March 2019, ECLI:NL:RBDHA:2019:1916.
Rb. Rotterdam 28 November 2019, ECLI:NL:RBROT:2019:9255.
Rb. Noord-Nederland 29 July 2020, ECLI:NL:RBNNE:2020:2633.
Rb. Overijssel 10 November 2020, ECLI:NL:RBOVE:2020:3694.

collected step and distance data such as collected by Van Zandwijk & Boztas (2019), emphasizing the applicability of the LR approach to the field of digital forensic science. The LR method will be demonstrated using a hypothetical case example in which a suspected arsonist's walking route is disputed. Lastly, this study will assess the performance of the LR method through investigating the discriminative power, calibration and accuracy of the method, and will examine the sensitivity of the LR to different case circumstances.

6.2 A fictive case example

We chose to design an LR method for the interpretation of distance data and not step count data as the former is more relevant to forensic casework. We consider the following case example to illustrate our likelihood ratio method for the interpretation of distance data in digital forensic science.

Scenario

A house, located at II in Fig. 6.1, was set on fire. Eyewitness X states that the fire started around 1:30 a.m. Suspect Y lives in the same neighborhood (location I). The suspect owns an **iPhone 7** equipped with the iPhone Health app which registers step count and traveled distances in time. The iPhone registered movement between 1:21 a.m. and 1:40 a.m. Specifically, 1450 steps and **1250 meters**. Eyewitness Z confirms that the suspect enters a bar (location III) around 1:40 a.m. The prosecution states that the suspect walked from location I to location III via location II (**1600 meters**). The suspect says he walked directly from location I to location III (**900 meters**).

Assumptions

The following assumptions are made:

- The suspect was nowhere to be seen on either route, except for the beginning and the end.
- Both routes are walkable within the time frame of 19 minutes (between 1:21 a.m. and 1:40 a.m.).
- The suspect's presence at location I and III is not disputed.
- The witness statements are considered reliable.
- The absence of step count at the time of fire might also be considered as evidence. However, this scenario only focuses on the distance registrations as evidence.

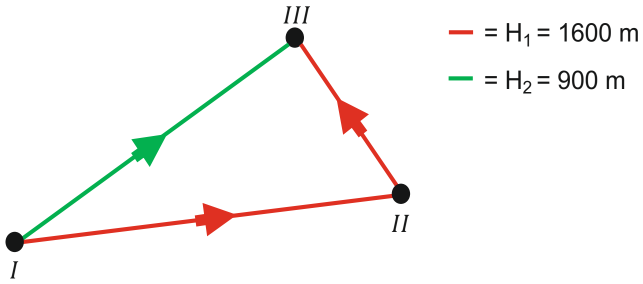


Figure 6.1: Schematic representation of location I (suspect's house), location II (location of fire), location III (the bar), and the walking routes in the scenario.

6.3 The LR framework

Rather than addressing the probability of a certain hypothesis, forensic scientists use their findings to express a degree of support for one hypothesis versus another. These hypothesis generally reflect the prosecution and defense view of events, H_1 and H_2 respectively. The degree of support is communicated through an LR, which measures the strength of (forensic) evidence. The LR is the ratio of two conditional probabilities. In forensic science, its form is commonly expressed as/in its discrete format (Aitken & Gold, 2013):

$$LR = \frac{P\left(E \mid H_1, I\right)}{P\left(E \mid H_2, I\right)} \tag{6.1}$$

with

E = the evidence

H_1, H_2 = the hypotheses

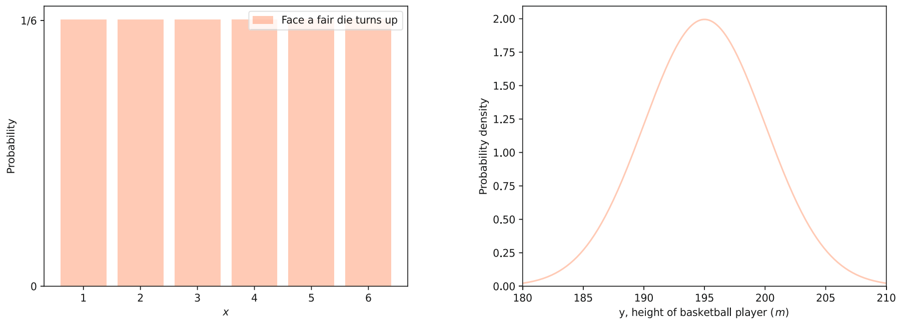
I = the relevant case information.

The LR in Eq. 6.1 signifies the evidential value of discrete evidence. A discrete variable only takes on a *countable* number of values. For example, consider evidence E as the random variable X, representing the face a fair die turns up. X is discrete and has possible values $x = 1, 2, 3, 4, 5, 6$ with probability $1/6$ each. The probability distribution of rolling a fair die can be presented using a histogram (Fig. 6.2a). The probability of each face equals the height of the corresponding bin in the histogram.

The evidence E can also be *measured* on a continuous scale. The evidence now takes on an uncountable and infinite number of possible values between a minimum and maximum value. For example, consider evidence E as the continuous variable Y, representing the height of a basketball player. The more accurate the measurements are, the more possible values for the random variable and the smaller the bins in the histogram. The histogram transitions into a probability density function (Fig. 6.2b). When the evidence E is a continuous variable, the LR is expressed as the ratio of two probability density functions evaluated at y, instead of two probabilities:

$$LR = \frac{f\left(y \mid H_1, I\right)}{f\left(y \mid H_2, I\right)} \tag{6.2}$$

with $f(y)$ being the probability density function of the continuous random variable Y.



(a) (b)

Figure 6.2: Histogram of the discrete variable X (a) and a probability density function of the continuous variable Y (b).

In our case example from Section 6.2, the covered distance of the suspect is disputed. The prosecution and defense hypotheses are:

$$H_1 : \text{Suspect Y walked from I to III via II } (1600\ m)$$
$$H_2 : \text{Suspect Y walked directly from I to III } (900\ m)$$

The evidence E is the distance registered by the iPhone Health app and equals 1250 m. The iPhone 7 is considered as case information I. The registered distance is obtained through measurements and is a continuous variable. Hence, the continuous LR in Eq. 6.2 conveys the strength of evidence from the iPhone Health app:

$$LR = \frac{f\,(\text{registered distance} \mid H_1, I)}{f\,(\text{registered distance} \mid H_2, I)} \tag{6.3}$$

The numerator and denominator specify the probability density functions of the registered distance when either H_1 or H_2 is considered as the true distance. Hence, two distinct probability density functions have to be evaluated to obtain an LR.

The LR connects the prior probability with the posterior probability of both hypotheses through Bayes' theorem[5]. An LR above 1 supports H_1 compared to H_2 and an LR below 1 supports H_2 compared to H_1; the more the LR deviates from 1, the more (relative) support it provides. An LR only provides relative support for one hypothesis versus the alternative. As such, there is not a single LR: every set of hypotheses, evidence and case information will result in a different LR. Moreover, the registered distance E on the iPhone might fit a third scenario better than the two scenarios in the hypotheses. For example, when the registered distance is 100 meters and the hypotheses include only 500 and 1000 meters. The forensic expert should mention this possibility in his report.

6.3.1 LR methods in forensic science

LRs can be expressed both verbally and numerically. Several computer-based methods have been designed to yield numerical LRs to assist the forensic expert's assessment in comparison problems. These methods are applied in multiple forensic disciplines such as gasoline comparison (Vergeer *et al.*, 2014), drug comparison (Bolck *et al.*, 2014), glass analysis (Van Es *et al.*, 2017), fingermark comparison (Leegwater *et al.*, 2017) and digital face comparison (Macarulla Rodriguez *et al.*, 2020). Comparison problems often evaluate whether two pieces of evidence share a common source (Ommen & Saunders, 2018).

Most of these methods are score-based and use feature comparison to define a measure of similarity or difference: a comparison score. This comparison score can be converted to an LR using several approaches (Aitken *et al.*, 2020; Brümmer & Du Preez, 2006; Haraksim *et al.*, 2017; Macarulla Rodriguez *et al.*, 2020; Morrison & Enzinger, 2018). One of these approaches is to model two probability density functions: the probability density functions of observing the score under both H_1 and H_2. Evaluating the two probability density functions

[5]The odds form of Bayes' rule multiplies the prior odds with the likelihood ratio to obtain the posterior odds.

at the comparison score (the numerator and denominator in Eq. 6.2 results in a numerical LR.

In essence, our method is very similar to LR methods previously described as it provides a probability model for two contrasting hypotheses. However, our (in general) relevant hypotheses specify a true value of a measurement (distance in meters) and not a trace's potential source:

H_1 : Person X walked route 1 (in m)
H_2 : Person X walked route 2 (in m)

In order to compute the numerator and denominator in Eq. 6.3, we will use an experimental dataset (see Section 6.4.1) in which both the true distance and the distance registered by the iPhone are known. This procedure is described in Section 6.4.2.

The hypotheses and case information are case-dependent and consequently, so is the model. Though the model is case-dependent, the method itself can be used for every case with compatible hypotheses - that is, if both hypotheses include a distance or a walking route that can be converted to a distance.

6.4 Our LR method

6.4.1 Data set

This research used a data set collected by Van Zandwijk & Boztas (2019). Van Zandwijk and Boztas conducted experiments to collect data on step count and distances in the iPhone Health application. Five individuals participated in the experiments. Each individual covered three different distances (A, B and C) with an iPhone 6(s) (iOS 10.2 and 10.3.3; grouped as one type: iPhone 6), iPhone 7 (iOS 10.3.1) and/or iPhone 8 (iOS 11.1.2) carried in five different locations (trouser pocket, two jacket pockets, backpack and hand). Each individual completed all three courses at least twice at walking pace and at least

	Walking pace					Running pace					
	Trouser pocket	Hand	Backpack	Jacket pocket high	Jacket pocket low	Trouser pocket	Hand	Backpack	Jacket pocket high	Jacket pocket low	Total
iPhone 6	$n = 113$	$n = 115$	$n = 107$	$n = 64$	$n = 63$	$n = 94$	$n = 99$	$n = 75$	$n = 54$	$n = 67$	$n = 851$
iPhone 7	$n = 65$	$n = 90$	$n = 88$	$n = 48$	$n = 43$	$n = 56$	$n = 71$	$n = 57$	$n = 40$	$n = 39$	$n = 597$
iPhone 8	$n = 34$	$n = 37$	$n = 25$	$n = 20$	$n = 15$	$n = 26$	$n = 34$	$n = 19$	$n = 19$	$n = 15$	$n = 244$
											$n = 1692$

Table 6.1: The number of observations per iPhone type, carrying location and walking pace. The total number of observations is 1692.

twice at running pace. One individual completing one course is considered a single run. Step count was also measured with a tally counter during all rounds, serving as the ground truth. Finally, the data set consisted of 1692 available data points (see Table 6.1). For a detailed description of the experimental set up, this paper refers to Van Zandwijk & Boztas (2019). The data set is available online[6].

6.4.2 Likelihood ratio method description

Fig. 6.3 illustrates the numerical LR method for the interpretation of distance data from the iPhone Health app. The aim of the method is to compute numerical LRs for hypotheses in which someone's walking distance is disputed. We chose to implement the method in the software Python (version 3.9) but other programming software are suitable too. The method consists of four steps: data selection and transformation, modeling, transforming, and LR computation.

Data selection and transformation

The data set was randomly split into a training set (60%, $n = 1015$) and a test set (40%, $n = 677$). The probability density functions were modeled using the training set and the LR method was validated (see Section 6.6 and Appendix C) using the test set. Two variables were selected from the training set: the

[6]https://github.com/NetherlandsForensicInstitute/iphone-health-app-data

6

registered distance X_r and the true distance X_t. These two original variables were transformed to a new statistic, the relative error e_r, expressed as:

$$e_r = \frac{x_r - x_t}{x_t} \tag{6.4}$$

with
e_r = the relative error between x_r, a registered distance, and x_t, the true distance.

The result was a set of relative errors for all data points in the training set. Van Zandwijk & Boztas (2019) showed that the relative error is a robust statistic as it is independent of the true distance. Hence, data from all three distances considered in Van Zandwijk & Boztas ($n = 1692$) could be used for modeling.

Modeling the probability density function of the relative error

Firstly, the method creates a histogram of the relative errors based on data from the three distances considered in Van Zandwijk & Boztas (2019). The method also considers type of iPhone, walking pace and carrying location as possible case information. If one of these variables is known as case information, the method will select the relevant data points from the training set and uses these data points to create the histogram. On the contrary, if no case information is available, all data points from the training set will be used to create the histogram.

Secondly, the continuous probability density function of the set of relative errors, $f(e_r)$, is estimated through kernel density estimation. Kernel density estimation produces a smooth continuous probability density function instead of a histogram with discrete bins. The bandwidth of the Gaussian Kernel represents the level of smoothing. The bandwidth is determined using both Scott's (1992) and Silverman's (1986) rules. The best bandwidth is identified through visual inspection of the Gaussian Kernel and is adjusted where

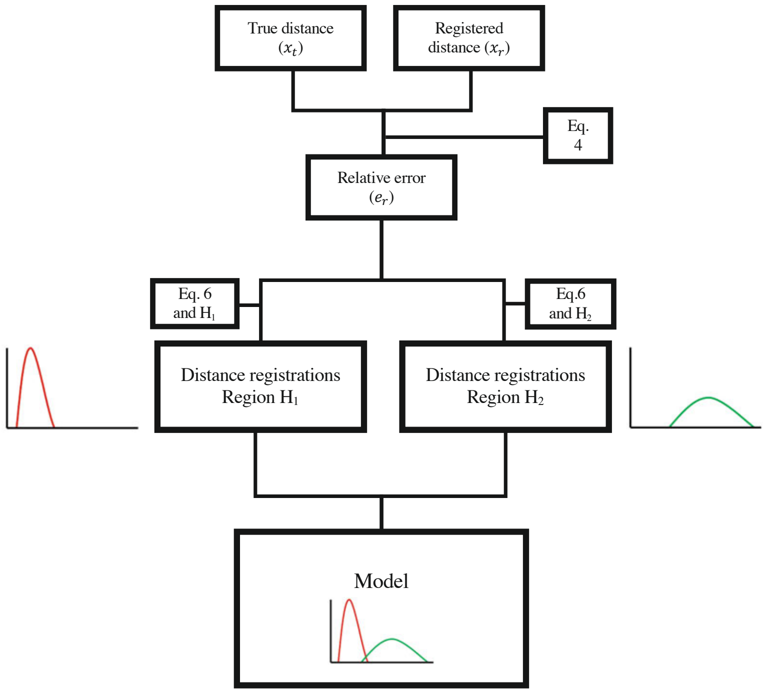
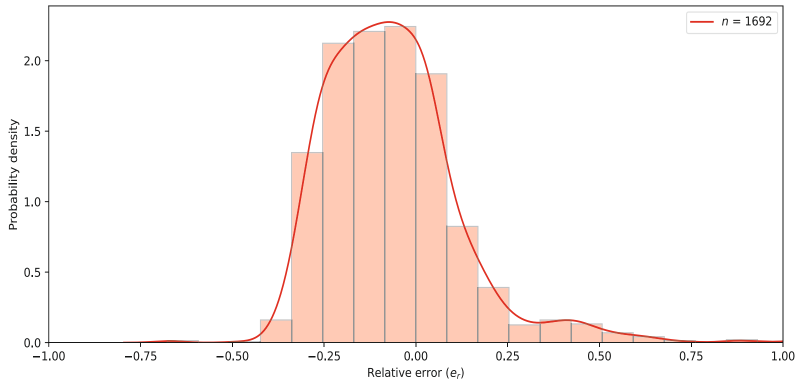


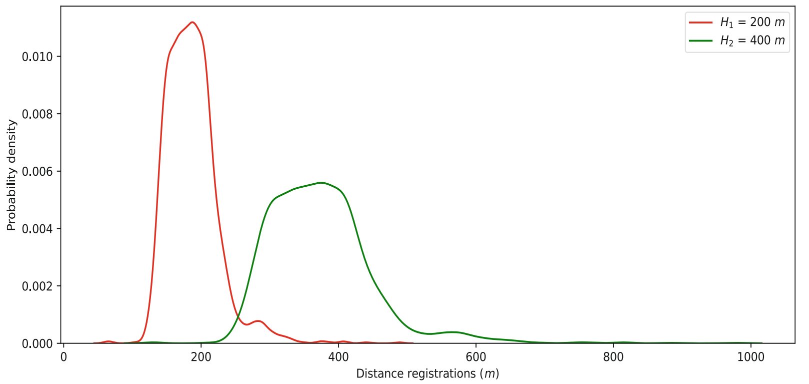
Figure 6.3: The LR method for the interpretation of distance data from the iPhone Health app.

needed. The intermediate result is one density function, based on data from all distances, from which both probabilities in the likelihood ratio can be obtained (see Fig. 6.4).

The probability density function in (Fig. 6.4) is in fact a weighted combination of different densities. This is because each combination of iPhone type, wearing location and walking pace yields a (slightly) different density. The frequencies of each combination in our experimental data determine the weights in the combined densities. In real cases, these experimental frequencies may not be suitable for the weighting factors, and instead personal probabilities may be used. We elaborate on this in the discussion section.



(a)



(b)

Figure 6.4: Probability density functions of (a) the relative error, $f(e_r)$, and (b) the distance registrations given H_1 : 200 m and H_2 : 400 m. (Eq. 6.6). No additional case information such as type of iPhone, walking pace or carrying location.

Transforming the relative error to the distance registrations given H_1 and H_2

Recall from Eq. 6.2 that the LR is the ratio of *two* probability functions of observing the evidence, E, given H_1 and H_2:

$$LR = \frac{f\left(E = x_r \mid x_t = H_1\right)}{f\left(E = x_r \mid x_t = H_2\right)} \tag{6.5}$$

Computation of the LR in 6.5 requires transformation of the probability density function of the relative error $f\left(e_r\right)$ to two probability density functions of the distance registrations given H_1 and H_2. Let e_r have the probability density function $f\left(e_r\right)$. The transformation of e_r to x_r follows from Eq. 6.4:

$$x_r = x_t(e_r + 1)$$

We can find the probability density function of the registered distance x_r using the rules for the density of transformed continuous random variables[7] as follows:

$$f_{x_t}(x_r) = f(e_r)\frac{d}{dx_r}\left(\frac{x_r - x_t}{x_t}\right) = f(e_r)/x_t$$

Using H_1 and H_2 as true known distances and the assumption that the distribution of the relative errors e_{r1} and e_{r2} also apply to other distances, we obtain two distributions: the probability density functions of observing a distance registration under both H_1 and H_2 respectively:

[7]If $Y = g\left(X\right)$ with $g\left(X\right)$ a strictly increasing function then $f_Y\left(y\right) = f_X\left(g^{-1}\left(y\right)\right)\frac{d}{dy}g^{-1}\left(y\right)$

$$
\begin{aligned}
f\left(x_r \mid H_1\right) &= f\left(e_{r1}\right)/H_1 \\
f\left(x_r \mid H_2\right) &= f\left(e_{r2}\right)/H_2
\end{aligned}
\tag{6.6}
$$

Note that, rather than being based on experimental data, these final two distributions result from scaling the relative error to two true known distances. Fig. 6.4 shows an example of two probability density functions of the distance registrations given $H_1 : 200\ m$ and $H_2 : 400\ m$.

LR computation

The LR is computed by evaluating the probability density of the registered distance in both the numerator and the denominator, using Eq. 6.5.

6.5 Results: a hypothetical case example

6.5.1 Model

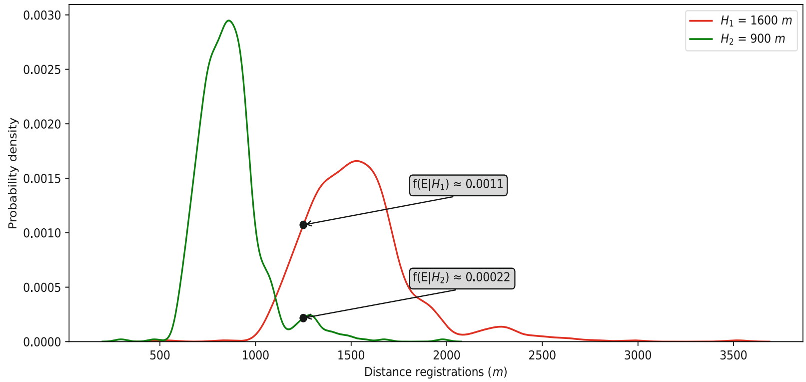


Figure 6.5: LR model for the case example with annotations of the probabilities in the likelihood ratio.

6

Fig. 6.5 shows the produced model ($n = 597$) for the hypotheses set $H_1 = 1600\ m$ and $H_2 = 900\ m$. Only the iPhone type (iPhone 7) is selected as case information (see Table 6.1 for the number of observations). This means that only distance registrations from the iPhone 7 were selected from the dataset to model the probability density functions. The red and green lines represent the probability density functions of the distance registrations given H_1 and H_2 respectively. The model shows little overlap between the probability distributions, reflecting good general discrimination properties. The overlap lies between 1000 m and 1500 m and the evidence of 1250 m is situated in this region. The overlap region is located in the tail of the probability distribution for H_2. Consequently, observing evidence in this region is almost always more likely given H_1 than given H_2.

Both distributions are slightly asymmetrical as they both have a long tail to the right. The smallest and largest observation given H_1 are approximately

1000 m and 2700 m. Most mass of the density is situated between 1250 m and 1750 m. The smallest and largest observation given H_2 are 500 m and 1500 m and most mass is situated between 500 m and 1000 m. Therefore, observing distance registrations smaller than the true distance is more probable than observing distance registrations larger than the true distance.

6.5.2 Likelihood ratio

The LR is derived by evaluating the ratio of the two probability density functions at $x_r = 1250\ m$ in Fig. 6.5. Recall from Eq. 6.5 that the LR is the ratio of the probability functions of observing the evidence E given H_1 and H_2:

$$LR = \frac{f\left(x_r = 1250\ m \mid x_t = 1600\ m, I\right)}{f\left(x_r = 1250\ m \mid x_t = 900\ m\ , I\right)} \approx 5 \tag{6.7}$$

The LR is approximately equal to 5 and offers weak support for H_1 rather than for H_2. To conclude, the case findings are slightly more probable[8] given H_1 than given H_2.

Both H_1 and H_2 contain a fixed distance, though it is very plausible to have (alternative) hypotheses that are ill-defined. Considering our case example, one possible set of (ill-defined) hypotheses (other than set H_1 and H_2 in section 2.2) are:

H_1 : Suspect Y walked from I to III via II (1600 m)
H_3 : Suspect Y walked directly from I to III (with a finite number of routes between I and III)

[8] Based on the verbal scale published in the *ENFSI Guideline for Evaluative Reporting in Forensic Science* (Willis *et al.*, 2015).

There will be no single LR for this set of hypotheses as H_3 consists of a range of likely distances traveled by suspect Y. Hence, there is no fixed distance to evaluate the probability density functions given H_3 with. One approach is to calculate an LR for each route (in meters) between I and III and to list these LRs in ascending order.

The method can yield larger LRs when the distance between H_1 and H_2 differs more and/or the evidence E is closer to one of the hypotheses. For instance, suppose that $H_1 = 2000\ m$, $H_2 = 900\ m$ and $E = 1750\ m$. Then:

$$LR = \frac{f\left(x_r = 1750\ m \mid x_t = 2000\ m, I\right))}{f\left(x_r = 1750\ m \mid x_t = 900\ m, I\right)} \approx 1147 \tag{6.8}$$

Fig. 6.6 illustrates the 10log LR (LLR) as function of the difference in distance between H_1 and H_2. The evidence E equals 1250 m, H_1 equals 1600 m and H_2 varies such that the difference between H_2 and H_1 varies between -1100 m and 400 m. The red dots represent the LLRs calculated with H_2 systematically varied. Left from 0 and right from 0 are the regions yielding LLRs for H_1 greater than H_2 and H_2 greater than H_1 respectively.

The LLR descents until H_2 roughly equals 1300 m and $x \approx -\ 300\ m$. The LLR starts to increase at $x \approx -300\ m$. The LLR is negative between $x \approx -400\ m$ and $x = 0\ m$, supporting H_2 rather than H_1. The green line represents the best polynomial fit LLR(x) for the red dots.

6

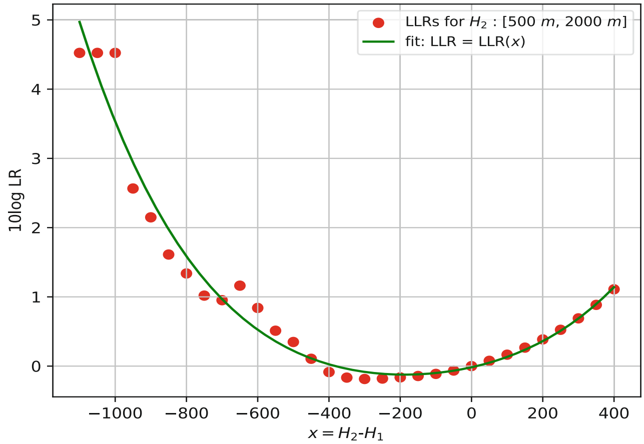


Figure 6.6: The 10log LR (LLR) as function of the difference in distance between H_2 and H_1. The red dots represent the calculated LLRs when H_2 is systematically varied and the green line shows the best polynomial fit for the LLR(x) : $5.2 \times 10^{-13}x^4 - 2.4 \times 10^{-9}x^3 + 2.4 \times 10^{-6}x^2 + 1.8 \times 10^{-3}x + 0.11$.

6.6 Performance measures

6.6.1 Methodology

We have established an LR method that computes numerical LRs for distance data from the iPhone Health app. We now need to assess the discriminative power of the method and whether the derived LRs are accurate and well-calibrated. Literature on the validation and calibration of forensic LR methods has increased over the past decade (Haraksim *et al.*, 2017; Leegwater *et al.*, 2017; Meuwly *et al.*, 2017; Morrison, 2011, 2021; Morrison & Enzinger, 2018; Morrison *et al.*, 2021; Ramos *et al.*, 2018; Vergeer *et al.*, 2021; Hannig & Iyer, 2021). Although many performance metrics and graphical representations are frequently reported as useful, a consensus on the best validation process to study the performance of a forensic LR method has yet to be established. We therefore chose the following measures to study the performance of the proposed LR method:

1. LR distributions, Tippett plots (Ramos *et al.*, 2013) and the rates of misleading evidence (Royall, 2000) are presented to study the discriminative power of the method.
2. The LRs are calibrated using a Pool Adjacent Violators algorithm (Brümmer & Du Preez, 2006; Fawcett & Niculescu-Mizil, 2007) and are visualized in PAV plots. The metric devPAV is used to study the difference between the pre- and post-calibrated LRs (Vergeer *et al.*, 2021).
3. ECE plots (Ramos *et al.*, 2018, 2013) are presented along with the log-likelihood ratio cost ($Cllr$) (Morrison, 2011, 2021; Ramos *et al.*, 2018) to measure the accuracy.
4. A sensitivity analysis is performed to study how sensitive the LR is to changes in case information.

The calibration (item 2) and accuracy (item 3) of the LR method are presented in Appendix C.
Most of the performance characteristics require data for which the true states (the ground truth) are known. In our data set, the only known true states are distance A (91.52 m), distance B (247.12 m) and distance C (450.93 m). Consequently, the method can only be validated for these three distances. If the hypotheses involve other distances, the discriminating power, accuracy and calibration may be affected. In that case, the forensic practitioner should preferably validate the case-specific model by conducting experiments on the distances stated in the case-specific hypotheses.

The models and LRs of three sets of hypotheses were validated and calibrated: set A,B, set A,C and set B,C. A, B and C represent the distances in the addressed hypotheses. For example, set A,B represents the hypotheses H_1 : Person X walked route 1 (91.52 m) and H_2 : Person X walked route 2 (247.12 m). The complete data set was randomly split into a training set

(60%, $n = 1015$) and a test set (40%, $n = 677$). The training set and test set were used to create and to validate the models respectively. For example, for set A,B, the entire training set (including all three distances A, B and C) was used to model the probability density functions given distance A and distance B (hereafter called "Model A,B"). Hence, both distances in the hypotheses are treated, although separately, as the ground truth. The test sets were forced to equal size and consisted of 292 trials with either a true distance A ($n = 146$) or B ($n = 146$). LRs were calculated using the model and the registered distances from the corresponding trials. This resulted in two sets of LRs: distance A LRs and distance B LRs. The sets of distance A LRs and distance B LRs considered distance A (or H_1) and distance B (or H_2) as the true distance respectively. All graphical representations and metrics were created using these sets of LRs. The Tippett plots, PAV plots, the devPAV, the ECE plots and the $Cllr$ (see Appendix A) were generated using a GitHub repository of the Netherlands Forensic Institute [9].

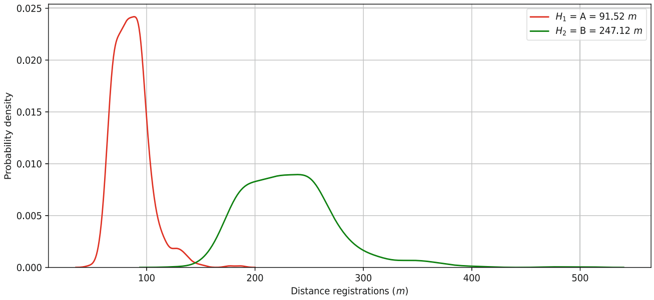
6.6.2 Results: discrimination

Models

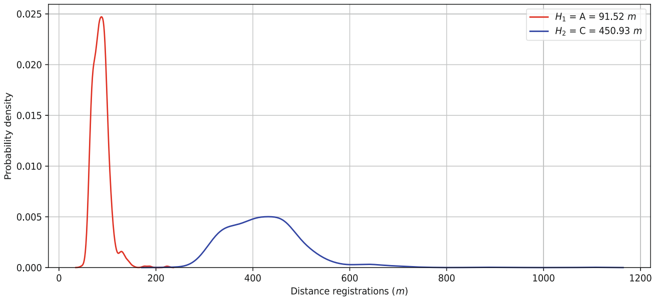
Fig. 6.7 shows the models for the hypotheses sets A,B (Fig. 6.7a), A,C (Fig. 6.7b) and B,C (Fig. 6.7c). Each model illustrates the probability density functions of the distance registrations given H_1 and H_2. The red, green and blue lines represent the probability density functions of A, B and C respectively. The models were used to derive the method-LRs for the hypotheses sets A,B, A,C and B,C by evaluating the ratio of the two probability distributions at a certain distance registration.

All distributions are unimodal and slightly skewed left. The distributions for H_1 are more narrow than the distributions for H_2. In general, most mass

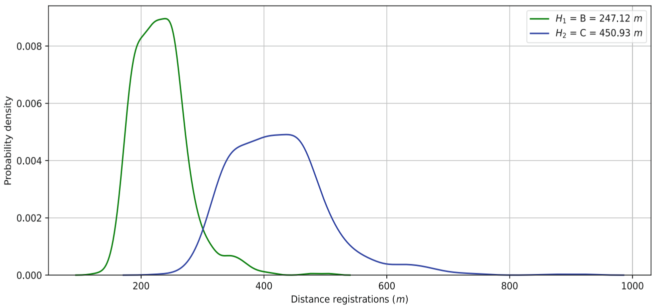
[9] https://github.com/NetherlandsForensicInstitute/lir



(a)



(b)



(c)

Figure 6.7: Probability density functions of the distance registrations in region H_1 and H_2 (LR model) for set A,B (a), set A,C (b) and set (B,C) (c).

6

of the distributions is concentrated around registrations smaller than the true distance. For example, most distance registrations in region $H_1 = \text{A} = 91.52$ m are concentrated below 91.52 meters. This indicates that the probability of observing smaller registrations than the true distances is larger than the probability of observing larger registrations.

Model A,B (Fig. 6.7) shows little overlap between both probability distributions, reflecting good discrimination properties. Model A,C (Fig. 6.7) shows even less overlap, reflecting almost perfect discrimination between the two hypotheses. The decrease in overlap between the two sets is expected as the difference in meters between distances A and C is greater than between distances A and B.

Model B,C (Fig. 6.7) exhibits the most overlap between the probability distributions. The overlap occurs between registrations of 250 m and 410 m. Hence, these distance registrations occur for both distance B and distance C. Distance registrations in this range may result in an LR supporting either H_1 or H_2.

LR distributions

The LRs for each set of hypotheses were derived for the test set using the models in Fig. 6.7. The probabilities of the distance registrations given the true distance A, B or C were evaluated using the model, resulting in six sets of LRs. For example, the ratio of the probability densities in model A,B resulted in two sets of LRs given that either A or B was true.

Fig. 6.8 shows the histograms of the resulting 10log LRs (LLRs) in region H_1 and H_2 for the hypotheses sets A,B (Fig. 6.8a), A,C (Fig. 6.8b) and B,C (Fig. 6.8c). The red, green and blue histograms represent the distance A LRs, the distance B LRs and the distance C LRs respectively. For set A,C, most LRs in region H_2 and H_1 go far beyond 10^{-50} and 10^{14} respectively. This can be explained through the high discriminative power of the model illustrated in Fig. 6.7. Due to the small overlap between the distributions of H_1 and H_2,

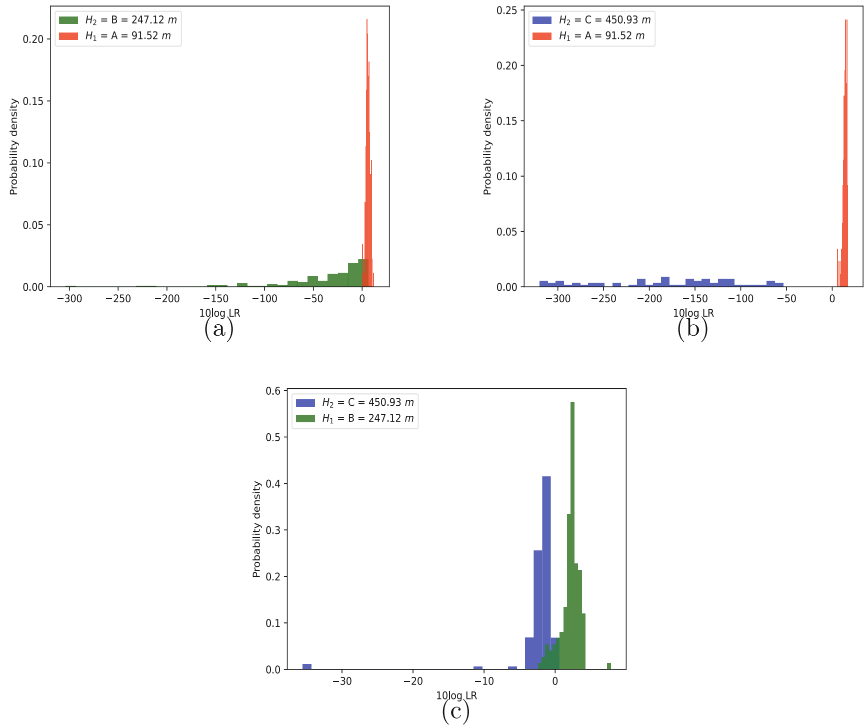


Figure 6.8: LR distributions for set A,B (a), set A,C (b) and set B,C (c).

a high probability in region H_1 corresponds to a small probability in region H_2 and vice versa. Consequently, the LR becomes either extremely large or extremely small.

Such LRs do not have any forensic meaning as there is little data in the corresponding tails of the probability distributions. This also applies for set A,B to some extent. In contrast, the overlap between the probability distributions for set B,C is larger and hence the LRs given H_1 ($Median = 232$) and H_2 ($Median = .021$) are far less extreme than for sets A,B and A,C. The mass of the distributions in region H_1 of set A,B and set A,C show a smaller spread and LRs given H_1 are weaker compared to LRs given H_2.

Tippett plots

Tippett plots are common graphical representations to visualize the discriminative power. Fig. 6.9 illustrates the Tippett plots for set A,B (6.9a), set A,C (6.9b) and set B,C (6.9c). The plots present the cumulative version of the LR distributions in Fig. 6.8, showing the proportion of LRs above or below a certain value. The greater the separation between the curves, the better the discrimination of the model. To conclude, just as illustrated in Fig. 6.8, the discriminative power of model A,B and A,C is higher than the discriminative power of model B,C.

Tippett plots allow us to study the percentage of misleading evidence. Evidence is misleading if LR > 1 when H_2 is true or LR < 1 when H_1 is true. If the performance of the system is good, the rates of misleading evidence will be small. Table C.1 shows an overview of misleading evidence for sets A,B; A,C and B,C. The percentage of misleading evidence is smaller than 10% for all sets; it is largest for set B,C and smallest for set A,C. The misleading evidence given H_2 for set A,B is very strong, including an LR of $\sim 10^6$. The majority of misleading evidence given H_1 and H_2 for set B,C include LRs between 0.01 and 1 and LRs between 1 and 4 respectively. To conclude, most of the misleading evidence for set B,C is weak.

Misleading evidence	A,B	A,C	B,C
LR $<$1 given H_1 (n = 146)	0.7%	0%	3%
LR $>$1 given H_2 (n = 146)	0%	0%	8%

Table 6.2: Percentages of misleading evidence for set A,B; A,C and B,C.

6.6.3 Results: secondary performance measures

Sensitivity analysis

Both calculations of the LR (Eq. 6.7 and Eq. 6.8) in our case example only

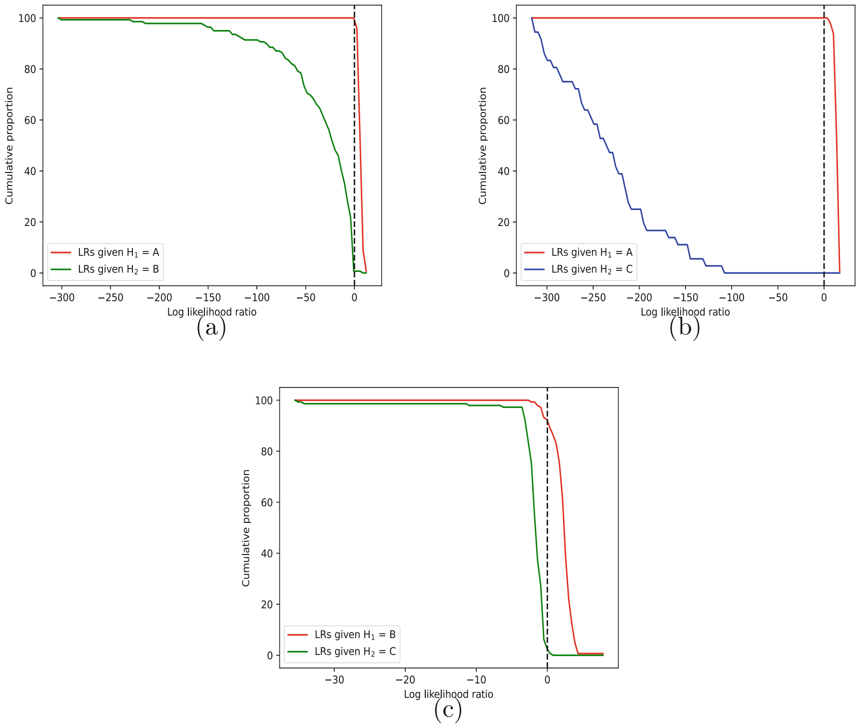


Figure 6.9: Tippett plots for set A,B (a), set A,C (b) and set B,C (c).

considered the iPhone type as known case information. However, the likelihood ratio differs as the case information alters. It is advantageous to know how sensitive the likelihood ratio is to changes in the case information. Table 6.3 shows the derived LRs for several combinations of known instances of the iPhone type, walking speed and carrying location. Most peculiarities exist in the column where running pace and trouser pocket are instantiated: depending on the iPhone type, the LR varies between 10^{-11} and 13. Thus the LR either moderately supports H_1 compared to H_2 or it extremely strongly supports H_2 compared to H_1. The LR differs the least among the iPhone types and the most for carrying location and walking speed. Knowing the carrying location of the iPhone can result in an LR in favor of either H_1 or H_2. This is in agreement with the conclusions in Van Zandwijk & Boztas (2019).

The column where walking pace and trouser pocket are instantiated shows extreme LR values. Such values may result from either perfect discrimination or poor kernel estimates of the tails of the probability density functions. This effect of the latter is enhanced with a lack of data. The more details we know, the less data we have and estimating the probability density function becomes more difficult. Therefore, the LR might not be well-calibrated and accurate. This emphasizes the importance of case-specific validation.

	Walking pace			Running pace			Unknown		
	Trouser pocket	Hand	Unknown	Trouser pocket	Hand	Unknown	Trouser pocket	Hand	Unknown
iPhone 6	10^{35} ($n = 113$)	86 ($n = 115$)	507 ($n = 462$)	13 ($n = 94$)	0.70 ($n = 99$)	5 ($n = 387$)	64 ($n = 207$)	0.85 ($n = 214$)	13 ($n = 851$)
iPhone 7	10^{21} ($n = 65$)	0.008 ($n = 90$)	23 ($n = 334$)	0.11 ($n = 56$)	0.62 ($n = 71$)	2 ($n = 263$)	11 ($n = 121$)	0.39 ($n = 161$)	5 ($n = 597$)
iPhone 8	10^{30} ($n = 34$)	0.12 ($n = 37$)	24 ($n = 131$)	10^{-11} ($n = 26$)	0.13 ($n = 34$)	0.77 ($n = 113$)	8 ($n = 60$)	0.24 ($n = 71$)	3 ($n = 244$)

Table 6.3: Likelihood ratios for $H_1 = 1600\ m$, $H_2 = 900\ m$, $E = 1250\ m$, and varied case circumstances I.

6.7 Discussion and conclusions

This paper has demonstrated an LR approach to the interpretation of distance data in digital forensic science by presenting an LR method for the comparison of alleged walking routes. A hypothetical case example illustrated the applicability of the LR method to actual casework but also showed its sensitivities and limitations. In our case example, the method yielded an LR of 5 for the hypotheses under consideration, yet it was also able to generate more extreme LRs. Varying the case information resulted in both small and extreme LRs. The extreme LRs were based on a small amount of data and their forensic meaning is therefore unreliable. This emphasizes the importance of continuous data collection and case-specific validation and calibration.

The validation process also illustrated that the method is able to yield extreme LRs. The method was validated using three hypotheses sets for which the true states were known: set A,B; A,C and B,C. All three models behaved well regarding discrimination and rates of misleading evidence. In fact, model A,C showed perfect discrimination between the two distances: there was no overlap between the two LR distributions and both the rates of misleading evidence. The high discriminative power resulted in extreme LRs given H_2 ($\sim 10^{-300}$). The model for set B,C exhibited most overlap between the LR distributions and therefore the LRs were far less extreme than for sets A,B and A,C. To summarize, this study has shown that it is possible to design an LR method that discriminates well between two distances. However, the difference in performance between the three models illustrates that the performance is highly case-dependent.

This paper advocates the use of the LR framework in digital forensic science. The method is a proof of concept rather than the optimal method for the comparison of walking routes. In real cases, the validation of the method requires additional work. Firstly, case-specific experiments need to be conducted with subjects representative for the suspect (or other persons of interest using the phone). If available, case-specific information such as walking pace or carrying location should be considered. Additionally, one may measure the performance of more distributional forms of the relative error. For instance, by choosing a parametric distribution rather than a kernel density estimate for the relative error.

Yet, information on wearing location and walking pace may be uncertain. There is no consensus in the literature on how to deal with unknown or uncertain case information such as carrying location and walking pace. In our case example, we derived the LR based on all data for the iPhone 7 ($n = 597$) and reported a single value for the LR. We ignored the absence of case information, although the sensitivity analysis shows the importance of dealing with its absence. There are two other solutions for handling the absence of case-specific information. One solution is to calculate the LR for all possible combinations of carrying location and walking pace separately and to report that the LR varies depending on the carrying location and walking pace (Sjerps *et al.*, 2016). A second solution (Taroni *et al.*, 2016) is to use personal probabilities of each carrying location/walking pace combination as a weighting factor to obtain a single probability density function, replacing the one in Fig. 6.4. This approach can be implemented in our method but this is left for future research.

The method was only validated for three sets of distances. If the hypotheses consist of other distances, the forensic practitioner should validate the case-specific model. The method and validation procedure is similar to those proposed in this study. The expert should i) collect reference data on the distances provided in the hypotheses, ii) generate the model (distribution fitting) using the method and iii) validate the model through assessing the performance measures.

During case-specific validation, one could investigate additional case information and add this as a new variable to the data set. Furthermore, the evidence mobile phone might contain useful historical data including walking distances with a known ground truth. For example, if the suspects walks a certain distance to his work each day. If such historical data is available, one can add these data as additional entries to the original data set. The discrepancies between the historical data and case data may also be used as additional case information.

More generally, the presented LR method may also be used in other applications of iPhone Health app data or for other data within and outside of the field of digital forensic science. Other possible applications of iPhone Health app data regard the number of stairs taken or the type of a user's physical activity such as walking or running (Van Zandwijk & Boztas, 2021). Other possible applications outside the field of digital forensic science regard, for instance, linking weapons to sharp force injuries through stab wound depth measurements. Here too, we measure a continuous feature (the depth of the wound) and compare it with two (or more) objects with known features (the length and width of the knife blade).

In conclusion, the proposed LR method is balanced, transparent and aids the forensic expert in their interpretation of mobile distance data in casework. Nevertheless, the performance of the method is highly case-dependent and it is unclear whether the data set is representative for other mobile phones or other case circumstances. Yet, the method and validation procedure are straightforward and can easily be repeated for different data sets from the iPhone Health app but also for other data in digital forensic science. By illustrating the LR method through the example of the comparison of walking routes, this study advocates the application of an LR approach to further relevant questions in digital forensic science.