STATISTICS IN MEDICINE
*Statist. Med.* 2008; **27**:1527–1538
Published online 20 August 2007 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.3031

# Improved accuracy when screening for human growth disorders by likelihood ratios

Stef van Buuren[1, 2, *, †]

[1]*TNO Quality of Life, P.O. Box 2215, 2301 CE Leiden, The Netherlands*
[2]*Department of Methodology and Statistics, FSS, University of Utrecht, Utrecht, The Netherlands*

SUMMARY

The standard deviation score (SDS) is a powerful tool for screening for growth-related problems. However, referral rules of the type 'if SDS($Y$)$<d$, then refer' (for some constant $d$) are not optimal for answering the question: 'Does this child with measurement $Y$ belong to the reference or to the diseased population?'. If the growth standard for the diseased population is known, then the likelihood ratio (LR) and the log-likelihood ratio (LLR) can be calculated for individual measurements. Rules of the type 'if LLR($Y$)$<e$, then refer' are uniformly the most powerful test for any constant $e$, implying that their receiver operating characteristic curves are above those for all other possible tests based on $Y$. As an empirical demonstration, both types of rules are applied to longitudinal growth data comparing a group with diagnosed Turner syndrome and a reference group from birth to 10 years of age. Conforming with theory, the LR rules were found to be superior to the SDS rules in terms of sensitivity and specificity. We conclude that the LR is the natural measure for two-group studies that can be easily calculated for individual measurements. The LR is firmly rooted within both statistical and decision theory and can be used to estimate the absolute probability of disease. Copyright © 2007 John Wiley & Sons, Ltd.

KEY WORDS: height; Turner's syndrome; screening; classification

## 1. INTRODUCTION

The standard deviation score (SDS) is a widely used measure for screening and referral in human growth [1]. The height SDS measures how the height $Y$ of a child deviates from the mean of the reference population of children of the same age and sex. The reference population $P_0$ typically consists of healthy children and is often represented by a height diagram, like the British [2] and Dutch [3]. The SDS values near zero indicate that a child's height is normal for age, whereas

*Correspondence to: Stef van Buuren, TNO Quality of Life, P.O. Box 2215, 2301 CE Leiden, The Netherlands.
†E-mail: Stef.vanBuuren@tno.nl

Contract/grant sponsor: Zorgonderzoek Nederland (ZON); contract/grant number: 2100.0050

*Received 25 September 2006*
*Accepted 26 June 2007*

values beyond the $\pm 2$ SD limits indicate that the child is either very tall or very short. The SDS is an extremely useful measure that corrects for age and sex differences and enables sensible comparisons across age.

In many situations, one would like to evaluate the probability of a specific disease given one or more height measurements $Y$. For example, a girl who is extremely short for age has an elevated chance of Turner syndrome (TS). Her height SDS measures how typical her height is within the general reference population $P_0$. It does not, however, tell how typical her height is within the diseased population $P_1$ of girls with the disorder. Also, the SDS alone will not predict how likely it is that the girl actually has the disease. When determining whether the child has a specified disease, more powerful diagnostic measures exist, such as the likelihood ratio (LR) and the log-likelihood ratio (LLR).

The LR can be understood by imagining the possibility of *two* SD scores for one $Y$: the conventional SD score $z_0$ for the reference population and an alternative SD score $z_1$ for the diseased population. Of course, $z_1$ can be calculated only if appropriate growth references are available for the disease of interest, but many such standards have been developed, e.g. for TS [4], Noonan syndrome [5], Prader–Willi syndrome [6], Silver–Russell syndrome [7], cri-du-chat syndrome [8], and so on. The LR combines the information provided by $z_0$ and $z_1$ to answer the question 'Does this child belong to the reference ($P_0$) or to the diseased ($P_1$) population?'. The LR is related to disease probability and can be used to estimate the costs of different screening scenarios. Similar to the SDS, the LR can be calculated for each individual data point.

This paper describes how the LR works in the context of height measurements to detect TS. The next section shows how to calculate the LR from two SDS scores, describes some of its properties, discusses the optimality of the LR, and explains how the LR improves upon screening rules that rely on SDS. Dutch data on TS are used to illustrate the principles.

## 2. METHOD

### 2.1. Likelihood ratio

The LR is a statistic for summarizing diagnostic accuracy. The LR measures how many times more likely patients with the disease are to have a particular result $Y$ than patients without the disease. Let $f_0$ and $f_1$ denote the density function for reference and diseased populations, respectively. The LR for a result $Y$ is defined by

$$\mathrm{LR}(Y) = \frac{f_1}{f_0} \tag{1}$$

Figure 1 illustrates the key concepts. Figure 1(a) contains two normal distributions. The distribution on the right-hand side corresponds to the variation in measurement $Y$ in the non-diseased population $P_0$. Here, $P_0$ is taken as the height reference standard of 6-year-old Dutch girls [3]. The distribution is normal with known mean $\mu_0 = 118.7$ cm and known standard deviation $\sigma_0 = 5.0$ cm. The distribution on the left-hand side represents how $Y$ varies in the diseased population $P_1$, 6-year-old girls with TS [4]. This is also normal distribution, with mean $\mu_1 = 104.5$ cm and standard deviation $\sigma_1 = 4.2$ cm. Thus, at the age of 6, girls with TS are on average $\delta = \mu_0 - \mu_1 = 118.7 - 104.5 = 14.2$ cm shorter. In the figure, the vertical axis is a relative frequency or a probability density.



DOI: 10.1002/sim

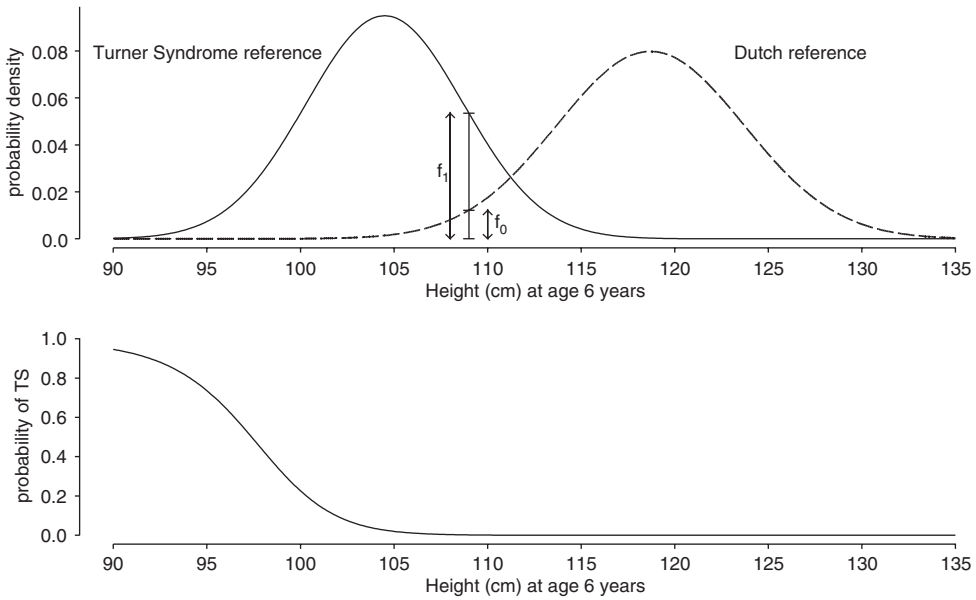


Figure 1. Illustration of the likelihood ratio for screening or growth-related disorders. (a) $f_0$ and $f_1$ are the density values at a height of 109 cm in the reference and syndrome populations. The likelihood ratio $\mathrm{LR} = f_1/f_0$. (b) Absolute probability of Turner syndrome as a function of height.

*2.1.1. Numerical example.* Suppose a 6-year-old girl has a height $Y = 109$ cm. This girl is short when compared with the Dutch reference (i.e. $z_0 = (Y - \mu_0)/\sigma_0 = (109 - 118.7)/5.0 = -1.94$ SD), but taller than girls with TS (i.e. $z_1 = (Y - \mu_1)/\sigma_1 = (109 - 104.5)/4.2 = +1.07$ SD). Figure 1(a) shows that at 109 cm the density in population $P_0$ is equal to $f_0 = f(Y, \mu_0, \sigma_0^2) = f(109, 118.7, 5.0^2) = 0.0122$, where $f(Y, \mu, \sigma^2)$ is the density at value $Y$ in a normal distribution with mean $\mu$ and variance $\sigma^2$. In population $P_1$, the density is equal to $f_1 = f(Y, \mu_1, \sigma_1^2) = f(109, 104.5, 4.2^2) = 0.0535$. The LR is thus equal to $\mathrm{LR}(Y) = f_1/f_0 = 0.0535/0.0122 = 4.4$. The interpretation of the $\mathrm{LR}(Y)$ is as follows: observing a height of $Y = 109$ cm is 4.4 times more likely in the Turner population $P_1$ than in the non-diseased population $P_0$.

The higher the value of $\mathrm{LR}(Y)$, the stronger the evidence for the presence of the disease. A value of $\mathrm{LR}(Y) = 1$ indicates that the measurement $Y$ is equally likely in populations $P_0$ and $P_1$. In Figure 1, this occurs at $Y = 111.2$ cm. In this case, a measurement $Y = 111.2$ cm does not discriminate between populations. Note that the LR is a relative measure and has nothing to do with the prevalence of the disease. If the pre-test probability is low, then even very large $\mathrm{LR}(Y)$ will not produce a large post-test probability of disease. Roberts [9] suggested that an $\mathrm{LR}(Y)$ of 10 provides strong evidence for the presence of the disease, although in many practical applications one should also take the prevalence of the disease into account.

$\mathrm{LR}(Y)$ can attain values between zero and infinity. It is generally more convenient to work with the natural logarithm of the LR, the LLR. One may calculate $\mathrm{LLR}(Y) = \ln(f_1/f_0) = \ln(f_1) - \ln(f_0)$. If the functions are $f_0 = f(Y, \mu_0, \sigma_0^2)$ and $f_1 = f(Y, \mu_1, \sigma_1^2)$, then $\mathrm{LLR}(Y)$ can be calculated from



DOI: 10.1002/sim

$z_0$ and $z_1$ as [9]

$$\mathrm{LLR}(Y) = 0.5(z_0^2 - z_1^2) + \ln(\sigma_0) - \ln(\sigma_1) \tag{2}$$

The first part compares the SD scores of $Y$ under populations $P_0$ and $P_1$. The second part compares the standard deviations. In the special case $z_0 = -z_1$ and $\sigma_0 = \sigma_1$, we find the points of indifference $\mathrm{LLR}(Y) = 0$ and $\mathrm{LR}(Y) = 1$, where $Y$ is equally likely under both $P_0$ and $P_1$.

*2.1.2. Numerical example.* In Figure 1, we find $z_0 = -1.94$ SD and $z_1 = +1.07$ SD at $Y = 109$ cm. Substituting these values gives $\mathrm{LLR}(Y) = (-1.94^2 - 1.07^2)/2 + \ln(5.0) - \ln(4.2) = 1.48$. The value of $\mathrm{LR}(Y)$ is equal to $\exp(1.48) = 4.4$, as before.

### *2.2. Properties of the LR*

The LR-statistic is well known in the statistical literature and possesses special properties.

*2.2.1. Optimality.* Suppose any cutoff $e$ is taken and that all subjects with score $\mathrm{LR}(Y) > e$ are classified as positive for disease. This test is the optimal test in the sense that its receiver operating characteristic (ROC) curve is everywhere above all other possible tests based on $Y$. This property is a direct consequence of the Neyman–Pearson fundamental lemma, which states that a test $\mathrm{LR}(Y) > e$ is the uniformly most powerful test [10, 11, Chapter 3]. Pepe [12, pp. 71, 269] provides an accessible description of the Neyman–Pearson theory in the context of medical tests.

*2.2.2. Prevalence and post-test disease probability.* The LR quantifies the knowledge about the presence of the disease that is gained through measurement $Y$. Define $D$ as a binary random variable, coded as $D = 1$ in the diseased population $P_1$, and as $D = 0$ in the reference population $P_0$. Let $P(D = 1)$ be the disease probability before knowing $Y$ (e.g. the prevalence), and let $P(D = 1|Y)$ be the disease probability given the measurement $Y$. Define the complement probabilities as $P(D = 0) = 1 - P(D = 1)$ and $P(D = 0|Y) = 1 - P(D = 1|Y)$. Using Bayes rule, the post-test odds $P(D = 1|Y)/P(D = 0|Y)$ of disease can be written as

$$\frac{P(D = 1|Y)}{P(D = 0|Y)} = \frac{P(Y|D = 1)P(D = 1)}{P(Y|D = 0)P(D = 0)} \tag{3}$$

where $P(Y|D = 0)$ and $P(Y|D = 1)$ are the probabilities of obtaining $Y$ in $P_0$ and $P_1$. Note that $\mathrm{LR}(Y) = P(Y|D = 1)/P(Y|D = 0)$; hence, multiplying the pre-test odds $P(D = 1)/P(D = 0)$ by $\mathrm{LR}(Y)$ produces the post-test odds. In practice, one could use (3) to calculate the posterior probability of disease by setting $P(D = 1)$ equal to the disease prevalence. If the pre-test disease probability $P(D = 1)$ is low and $\mathrm{LR}(Y)$ is not huge, (3) is approximated by $P(D = 1|Y) \approx \mathrm{LR}(Y) \times P(D = 1)$.

*2.2.3. Numerical example.* For TS, the prevalence in the general population is low, about 1:2500 girls. The probability of TS for a 6-year-old girl with a height of 109 cm is thus approximated by $P(D = 1) \approx 4.4 \times 1/2500 = 0.00176$. Figure 1(b) illustrates how $P(D = 1)$ varies with height. It appears that $P(D = 1)$ is sizeable only for the extremely short girls. At 107.7 cm, $\mathrm{LR}(Y) = 10$, whereas the $P(\mathrm{TS})$ is only 0.004. Observe that, despite a 'high' LR of 10, 99.6 per cent of the girls with a height of 107.7 cm will *not* have TS.



DOI: 10.1002/sim

*2.2.4. Binormal ROC curves.* Suppose that $Y$ is normally distributed as $Y \sim \mathrm{N}(\mu_0, \sigma_0^2)$ and $Y \sim \mathrm{N}(\mu_1, \sigma_1^2)$ in populations $P_0$ and $P_1$, respectively. For some threshold value $c$, let the *false positive fraction* $\mathrm{FPF}(c) = P(Y<c|D=0) = \Phi((c-\mu_0)/\sigma_0)$ and the *true positive fraction* $\mathrm{TPF}(c) = P(Y<c|D=1) = \Phi((c-\mu_1)/\sigma_1)$, where $\Phi$ is the cumulative normal distribution function and the *sensitivity* of a test $Y<c$ is equal to $\mathrm{TPF}(c)$, whereas the *specificity* $= 1-\mathrm{FPF}(c)$. The ROC plot displays $\mathrm{TPF}(c)$ against the $\mathrm{FPF}(c)$ for an increasing sequence of thresholds $c$. For binormal references, $\mathrm{TPF}(c)$ and $\mathrm{FPF}(c)$ are related as follows:

$$\mathrm{TPF}(c) = \Phi(a + b\Phi^{-1}(\mathrm{FPF}(c)) \tag{4}$$

where $a = (\mu_0 - \mu_1)/\sigma_1$ and $b = \sigma_0/\sigma_1$ [12, p. 82]. One could use this result to calculate sensitivity at a specific cutoff. For example, for a test $Y<109$ cm, we find $\mathrm{FPF}(109) = P(Y<109|P=0) = \Phi(-1.94) = 0.026$. The sensitivity is equal to $\mathrm{TPF}(109) = \Phi(+1.07) = 0.858$. Alternatively, applying (4) at a FPF(109), we find $\mathrm{TPF}(109) = \Phi((118.7 - 104.5)/4.2 + (5.0/4.2) \times \Phi^{-1}(0.026)) = 0.858$.

### *2.3. SDS versus LR rules*

The possible classification rules for a measurement $Y$ are $Y<c$, $\mathrm{SDS}(Y)<d$ and $\mathrm{LLR}(Y)>e$ for some $c$, $d$, or $e$. If we know the distribution of $Y$, we can choose $c$, $d$, or $e$ and calculate the other two cutoff values. As long as the relation between $Y$ and $\mathrm{LLR}(Y)$ is monotonic, it does not matter which rule we take.

Figure 2 shows how potential cutoff points $d$ and $e$ vary with height for 6-year-old Dutch girls. Both SDS and LLR are monotonically related to height. The relationship between SDS and

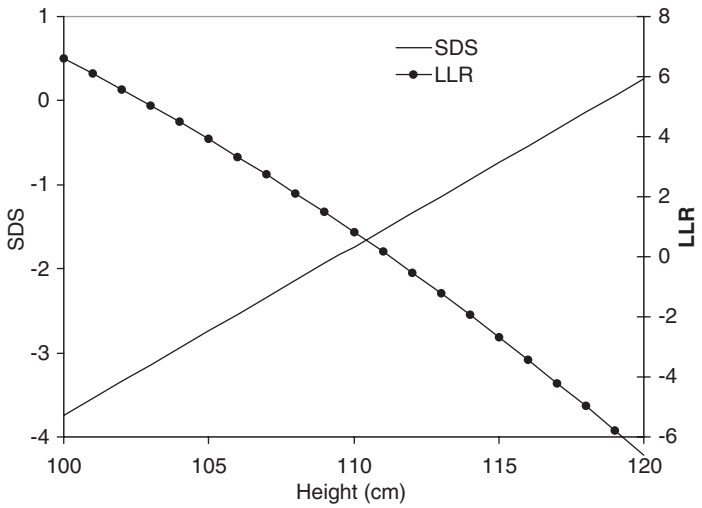


Figure 2. SDS (left axis) and LLR (right axis) values as a function of height for Dutch 6-year-old girls. Both SDS and LLR are monotonically related to height (SDS linear, LLR quadratic).



DOI: 10.1002/sim

height is linear, whereas that between LLR and height is quadratic. For a 6-year-old girl, the rules $Y<109$ cm, $\mathrm{SDS}(Y)<-1.94$ SD and $\mathrm{LLR}(Y)>1.48$ are all equivalent and have the same (local) sensitivity and specificity.

Complications arise if the distribution of $Y$ varies with age or sex. In that case, fixing $c$, $d$, or $e$ to some value generally corresponds to very different test strategies. A rather absurd example is the test $Y<109$ cm at age 2 (instead of age 6). The test would be true for everyone and thus have 100 per cent sensitivity and 0 per cent specificity. This leads to the question of which rule should be taken. The Neyman–Pearson lemma states that the rule $\mathrm{LLR}(Y)>e$ is superior to any other rules based on $Y$. Let us now look at the interpretation of the rules $\mathrm{SDS}(Y)<d$ and $\mathrm{LLR}(Y)>e$ in more detail.

Current height screening rules generally assume equal specificity across age. For example, height screening rules such as 'if $\mathrm{SDS}(Y)<d$, then refer child' imply identical specificity across age. The rule does not take any aspects of the diseased group into account. Note that the sensitivity of the rule will critically depend on 'how far' the diseased population is from the reference. If this distance varies with age, then the sensitivity of the $\mathrm{SDS}(Y)<d$ rule will generally also depend on age. For the same reason, the $\mathrm{LLR}(Y)$ value corresponding to the test $\mathrm{SDS}(Y)<d$ at some age will not be constant across age.

An alternative is a rule of the type 'if $\mathrm{LLR}(Y)>e$, then refer child'. This rule is expected to be more efficient since the strength of the evidence (as measured by the LLR) that the child is a case will be identical across age. In rules based on $\mathrm{LLR}(Y)$, both sensitivity and specificity depend on age, whereas $\mathrm{LLR}(Y)$ is constant.

*2.3.1. Numerical example.* At two years of age, the parameters are $\mu_0=87.5$, $\mu_1=80.6$, $\sigma_0=3.2$ and $\sigma_1=3.1$. The cutoff point at age 2 for which $\mathrm{LLR}(Y)=1.48$ must solve for $c$ in $f(c, 80.6, 3.1)-f(c, 87.5, 3.2)=1.48$. The solution is $c=81.9$ cm, so $d=(81.9-87.5)/3.2=-1.75$. Thus, at age 2, the tests $Y<81.9$ cm, $\mathrm{SDS}(Y)<-1.75$ and $\mathrm{LLR}(Y)>1.48$ are equivalent. The local specificity of this test is 0.960 and the local sensitivity is 0.662, both of which are inferior to the test $\mathrm{LLR}(Y)>1.48$ at age 6 (specificity: 0.974, sensitivity: 0.858). This reflects that it is more difficult to find cases at age 2 than at age 6. In comparison, the specificity of the test $\mathrm{SDS}(Y)<-1.75$ is constant at 0.960 and has a local sensitivity of 0.903 at age 6. Thus, the tests $\mathrm{SDS}(Y)<d$ and $\mathrm{LLR}(Y)>e$ are different at different ages.

The Neyman–Pearson lemma implies that the rule $\mathrm{LLR}(Y)>e$ is preferable if there are two or more groups. To see how this works, suppose the sample consists of two groups, A and B, of sizes $n_\mathrm{A}$ and $n_\mathrm{B}$, respectively. If $\mathrm{TPF_A}$ and $\mathrm{TPF_B}$ are the sensitivities per group, then the sensitivity in the groups combined is equal to $\mathrm{TPF}=(n_\mathrm{A}\mathrm{TPF_A}+n_\mathrm{B}\mathrm{TPF_B})/(n_\mathrm{A}+n_\mathrm{B})$. Similarly, $\mathrm{FPF}=(n_\mathrm{A}\mathrm{FPF_A}+n_\mathrm{B}\mathrm{FPF}_B)/(n_\mathrm{A}+n_\mathrm{B})$. Suppose that we apply two rules on the sample: $\mathrm{SDS}<d$ and $\mathrm{LLR}(Y)>e$. For each $d$, we can always choose $e$ such that FPF in the sample is equal under both rules. The Neyman–Pearson lemma implies that $\mathrm{TPF(LLR)}\geqslant\mathrm{TPF(SDS)}$. Alternatively, we could equate sensitivity and find $\mathrm{FPF(LLR)}\leqslant\mathrm{FPF(SDS)}$.

*2.3.2. Numerical example.* In the above example, the specificity of rule $\mathrm{SDS}(Y)<-1.75$ is 0.960, with sensitivities 0.662 (2 years) and 0.903 (6 years). If $n_A=n_B$, then $\mathrm{TPF(SDS)}=0.7824$. The specificity of rule $\mathrm{LLR}(Y)>1.26$ is 0.960, with sensitivities 0.699 (2 years) and 0.876 (6 years). In this case, $\mathrm{TPF(LRR)}=0.7872$; hence, indeed $\mathrm{TPF(LRR)}>\mathrm{TPF(SDS)}$. Note that the gain in sensitivity is rather small here, but, as we will see, in practice the difference can be much larger.



DOI: 10.1002/sim

### 2.4. Materials

Two samples of data were analysed. Longitudinal height data from 777 girls with untreated TS were collected from several sources. A reference sample of longitudinal height data was retrospectively obtained for a cohort of 489 girls born in 1989 and 1990 in Landgraaf. More details on these data can be found elsewhere [13].

### 2.5. Statistical methods

SDS and LLR values were graphically compared in both reference and disease groups by plotting the individual data points against age. Two age-dependent screening rules were formulated, one using SDS and the other using the LLR. The rules were operated on the child level. Both rules refer a girl if at least one of her height measurements is beyond the stated cutoff point. The discriminatory power of both rules was compared through case–control simulation [13]. The cutoff points $d$ and $e$ were varied continuously in the range $-10$ and $+20$. The sensitivity and specificity were calculated at each cutoff value. The results are presented as two curves in a ROC plot.

## 3. RESULTS

Figure 3 plots the individual SDS values according to age, for both the reference girls and the TS girls. The average SDS of the reference girls is slightly lower than the national reference.

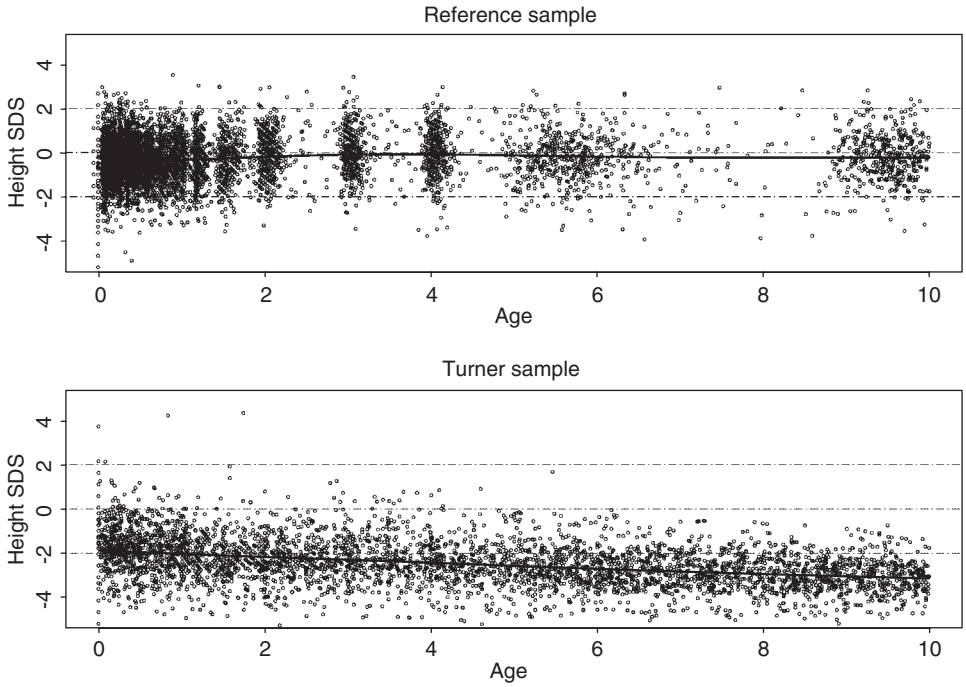


Figure 3. Individual SDS values plotted against age for the control and Turner samples.



DOI: 10.1002/sim

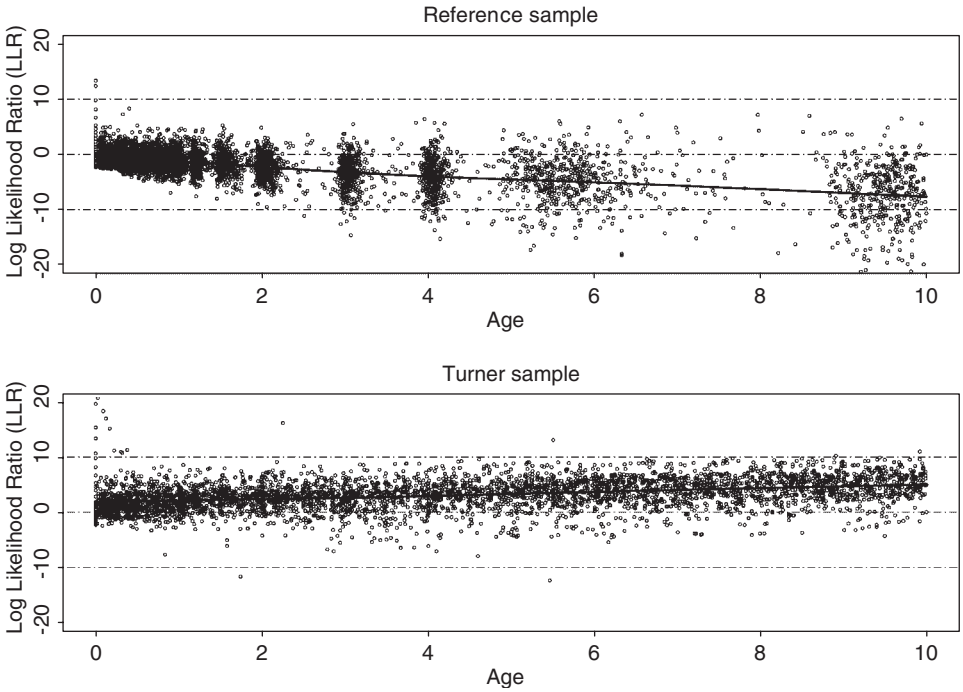


Figure 4. Individual LLR values plotted against age for the control and Turner samples.

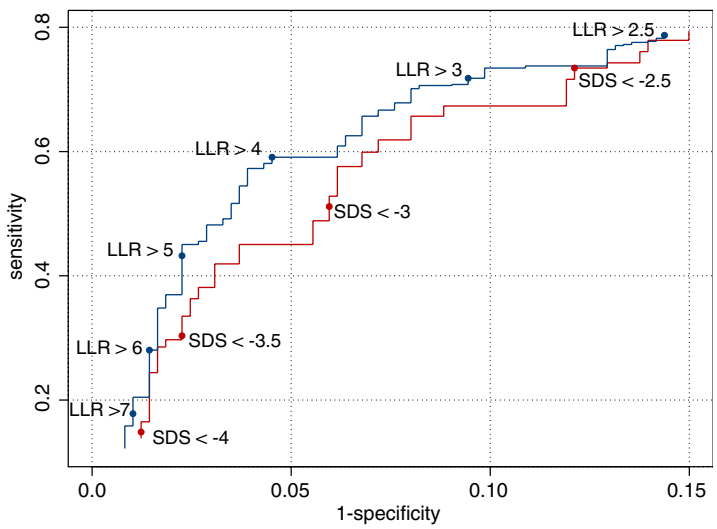


Figure 5. ROC plot for the LLR and SDS screening rules. All ages (0–10 years) are combined. A child is referred if it meets a referral criterion on at least one occasion. Sensitivity and specificity are properties of the referral procedure covering ages 0–10 years. The plot shows that rules based on the LLR are superior to those based on SDS for the detection of TS.



DOI: 10.1002/sim

This finding was expected because these girls are living in a part of the Netherlands were children and are somewhat shorter than the general Dutch population [3]. By contrast, the mean SDS of the Turner group has a clear downward trend. The height gap between the reference and TS girls grows with age. This pattern suggests that screening rules of the type $\mathrm{SDS}(Y)<d$ may work well for the older girls, but are less useful during infancy and early childhood.

Figure 4 presents the same data on the LLR scale. As before, the probability of TS increases with age in the TS group. More interestingly, the mean of the control group has a downward trend. This means that, as age increases, it becomes less probable that these girls will have TS. Note that this information is not present in Figure 3.

Figure 5 shows the ROC plot with the SDS and LLR rules for all ages combined. Figure 4 shows that the discriminatory power of the LLR rules is superior to that of the SDS rule. The differences are especially large in the area where the specificity is around 95 per cent, the area that is generally of most interest for screening purposes.

## 4. DISCUSSION

This paper introduces the likelihood ratio (LR) and the log likelihood ratio (LLR) in the context of screening for a particular growth-related disease. The LLR measures how much more likely the observation is in the diseased than in the reference group. The LLR is easy to calculate for individual measurements, is firmly rooted within both statistical and decision theory, and can be used in screening settings to estimate the absolute probability of the disease, given the pre-test probability. We provide both theoretical and empirical evidence that screening rules based on the LLR discriminate better than rules based on the SDS.

The LR is the natural companion of the SDS. Until the 1930s, statistical inference was mainly concerned with single (null) distributions. The LR was introduced by Fisher [14] and was given its modern interpretation by Neyman and Pearson [15]. The LR naturally arises out of their notion that statistical inference involves both a null and an alternative hypothesis. Although the single group approach—and the associated SDS and percentile scores—still dominates the field of human growth, attention seems to be shifting towards the two-group perspective [16]. The LR is the natural measure for two-group studies.

The improved performance of the LR rule occurs if the overlap between the non-diseased ($P_0$) and diseased ($P_1$) populations varies with age. At a given age, one can always choose $c$, $d$ or $e$ such that the three types of classification rules $Y<c$, $\mathrm{SDS}(Y)<d$ and $\mathrm{LLR}(Y)>e$ have identical sensitivity and specificity. However, the relationship between $c$, $d$, and $e$ depends on the amount of overlap of the two distributions, and thus on age. One could fix $c$, $d$ or $e$ and calculate the other two. The Neyman–Pearson fundamental lemma tells us that fixing $e$ across age is preferable to fixing $c$ or $d$ because that provides us with the uniformly most powerful test.

Since $c$, $d$ and $e$ are related, in practice one can perform all screening in the $Y$-metric, i.e. by applying the $Y<c$ rule. In this case, the cutoff values $c$ need to vary by age and sex. If there is no diseased population, we may calculate time-varying values $c$ such that $d$ is constant across age and sex, which is equivalent to assuming constant specificity. If we screen for a target disease, then the uniformly most powerful test corresponds to values for $c$ such that $e$ is constant across age and sex, which fixes the amount of evidence as measured by the LR. Thus, in practice, there is no need to calculate either the SDS or the LLR for a given child. We only need access to a table of appropriate age- and sex-dependent referral values $c$, e.g. in cm or inches. Without doubt,



DOI: 10.1002/sim

Table I. The results of calculating the referral height in cm ($c$) for rules SDS$< -2.5$ and LLR$>3$.

| | Dutch reference | | Turner reference | | SDS$<-2.5$ | | | | LLR$>3$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Age | Mean | SD | Mean | SD | $c$ | $d$ | $e$ | Sensitivity | Specificity | $c$ | $d$ | $e$ | Sensitivity | Specificity |
| 0 | 50.9 | 1.9 | 47.6 | 2.5 | 46.2 | −2.50 | 2.68 | 0.28 | 0.994 | 45.9 | −2.65 | 3.00 | 0.24 | 0.996 |
| 0.5 | 66.4 | 2.3 | 62.2 | 2.6 | 60.7 | −2.50 | 2.82 | 0.28 | 0.994 | 60.4 | −2.59 | 3.00 | 0.25 | 0.995 |
| 1 | 75.1 | 2.6 | 69.9 | 2.8 | 68.6 | −2.50 | 2.94 | 0.32 | 0.994 | 68.5 | −2.53 | 3.00 | 0.31 | 0.994 |
| 1.5 | 82.1 | 2.9 | 76.1 | 2.9 | 74.9 | −2.50 | 3.03 | 0.33 | 0.994 | 74.9 | −2.48 | 3.00 | 0.34 | 0.994 |
| 2 | 87.5 | 3.2 | 80.6 | 3.1 | 79.5 | −2.50 | 3.09 | 0.36 | 0.994 | 79.6 | −2.46 | 3.00 | 0.38 | 0.993 |
| 3 | 96.7 | 3.7 | 87.6 | 3.4 | 87.5 | −2.50 | 3.21 | 0.48 | 0.994 | 87.8 | −2.42 | 3.00 | 0.52 | 0.992 |
| 4 | 104.5 | 4.1 | 93.7 | 3.7 | 94.3 | −2.50 | 3.22 | 0.56 | 0.994 | 94.6 | −2.42 | 3.00 | 0.59 | 0.992 |
| 5 | 111.8 | 4.6 | 99.3 | 3.9 | 100.3 | −2.50 | 3.26 | 0.60 | 0.994 | 100.7 | −2.41 | 3.00 | 0.64 | 0.992 |
| 6 | 118.7 | 5.0 | 104.5 | 4.2 | 106.2 | −2.50 | 3.22 | 0.66 | 0.994 | 106.6 | −2.43 | 3.00 | 0.69 | 0.992 |
| 7 | 125.2 | 5.4 | 109.5 | 4.4 | 111.7 | −2.50 | 3.20 | 0.69 | 0.994 | 112.1 | −2.43 | 3.00 | 0.72 | 0.993 |
| 8 | 131.5 | 5.7 | 114.1 | 4.6 | 117.3 | −2.50 | 3.10 | 0.75 | 0.994 | 117.4 | −2.47 | 3.00 | 0.77 | 0.993 |
| 9 | 137.5 | 6.1 | 118.5 | 4.8 | 122.3 | −2.50 | 3.06 | 0.78 | 0.994 | 122.4 | −2.48 | 3.00 | 0.79 | 0.993 |
| 10 | 143.3 | 6.4 | 122.5 | 5.0 | 127.3 | −2.50 | 2.91 | 0.83 | 0.994 | 127.1 | −2.52 | 3.00 | 0.82 | 0.994 |

*Note*: Values labelled $c$ can be used as age-dependent referral values based on the cm scale.



DOI: 10.1002/sim

the ability to perform optimal screening tests by just comparing a raw measurement $Y$ with a tabulated value $c$ is a tremendous practical benefit.

As an illustration of the above point, Table I provides age-specific cutoff points for rules SDS$<$2.5 and LLR$>$3 calculated from the Dutch and Turner references. The column labelled '$c$' contains the cutoff values for height (cm). The SDS rule fixes $d$ and the specificity over age, and allows $e$ to vary. The LLR rule fixes $e$ and allows $d$ to vary. Note that the differences between the LLR and SDS rules in terms of absolute height or the SDS scale are generally not large. For younger children the LLR rule is stricter than the SDS rule. The differences between the Dutch and Turner references are relatively small for young children; hence, the LLR rule becomes more conservative by increasing specificity. The situation is reversed for older ages.

Screening for more than one disease at the same time is often desired. Michael Hermanussen (personal communication) made an interesting suggestion for calculating multiple absolute probabilities for different diseases. This would take the prevalence of disorder into account. The list of disease-specific probabilities is presented as evidence to the clinician. There is yet no practical experience of such methods. To be realistic, the potential for discriminating between different disorders that all lead up to short stature may be limited. On the other hand, the fact that this approach takes the prevalence into account may turn out to be a great asset. One of the reviewers pointed out a connection with the work of Spiegelhalter and Knill-Jones [17], who discussed a Bayesian decision tree for multiple possible diagnoses in which the 'weight of evidence' is a LLR. Application of this Bayesian decision tree approach to multiple diagnoses in human growth is a promising line of further development.

The LR as described takes only height into account and ignores any other clinical symptoms that might be present. It is, in principle, possible to combine several independent pieces of evidence into one LR. This also provides a way to include the child's own measurements in longitudinal settings. Such options would be useful extensions of the present methodology.

## ACKNOWLEDGEMENTS

I thank Ine Bonnemaijer for her cooperation in obtaining the reference group data, and Anita Hokken-Koelega, Gladys Zandwijken, Sabine de Muinck Keizer-Schrama and Ciska Rongen-Westerlaken for their cooperation in obtaining the Turner data. Jan Maarten Wit provided useful comments on a previous version of the manuscript. I thank the associate editor and two anonymous referees for their insightful comments, which helped in improving the presentation. This research was financially supported by grant number 2100.0050 from Zorgonderzoek Nederland (ZON) entitled 'Objectivering van verwijscriteria bij biometrisch onderzoek in de jeugdgezondheidszorg: Pilot Turner Syndroom'. The funding source had no involvement in the work.

## REFERENCES

1. Cole TJ. Do growth chart centiles need a face lift? *British Medical Journal* 1994; **308**(6929):641–642.
2. Freeman JV, Cole TJ, Chinn S, Jones PRM, White EM, Preece MA. Cross sectional stature and weight reference curves for the UK, 1990. *Archives of Disease in Childhood* 1995; **73**:17–24.
3. Fredriks AM, van Buuren S, Burgmeijer RJ, Meulmeester JF, Beuker RJ, Brugman E, Roede MJ, Verloove-Vanhorick SP, Wit JM. Continuing positive secular growth change in The Netherlands 1955–1997. *Pediatric Research* 2000; **47**:316–323.
4. Rongen-Westerlaken C, Corel L, van den Broeck J, Massa G, Karlberg J, Albertsson-Wikland K, Naeraa RW, Wit JM. Reference values for height, height velocity and weight in Turner's syndrome. Swedish Study Group for GH treatment. *Acta Paediatrica* 1997; **86**:937–942.



DOI: 10.1002/sim

5. Ranke MB, Heidemann P, Knupfer C, Enders H, Schmaltz AA, Bierich JR. Noonan syndrome: growth and clinical manifestations in 144 cases. *European Journal of Pediatrics* 1988; **148**:200–227.
6. Grauer ML, Wollmann HA, Schulz V, Ranke MB. Reference values for height and weight in Prader–Willi syndrome based on 315 patients. *Hormone Research* 1997; **48**(Suppl. 2):54.
7. Wollman HA, Kirchner T, Enders H, Preece MA, Ranke MB. Growth and symptoms in Silver–Russell syndrome: review on the basis of 386 patients. *European Journal of Pediatrics* 1995; **154**:958–968.
8. Marinescu RC, Mainardi PC, Collins MR, Kouahou M, Coucourde G, Pastore G, Eaton-Evans J, Overhauser J. Growth charts for cri-du-chat syndrome: an international collaborative study. *American Journal of Medical Genetics* 2000; **94**(2):153–162.
9. Roberts RS. Likelihood ratio with diagnostic tests. In *Encyclopedia of Biostatistics*, Armitage P, Colton T (eds). Wiley: NY, 1998; 2248–2253.
10. Neyman J, Pearson ED. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London*, *Series A* 1933; **231**:289–337.
11. Lehmann EL. *Testing Statistical Hypotheses* (2nd edn). Wiley: New York, 1986.
12. Pepe MS. *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press: Oxford, 2003.
13. van Buuren S, van Dommelen P, Zandwijken GR, Grote FK, Wit JM, Verkerk PH. Towards evidence based referral criteria for growth monitoring. *Archives of Disease in Childhood* 2004; **89**(4):336–341.
14. Fisher RA. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London*, *Series A* 1922; **222**:309–368.
15. Neyman J, Pearson ED. On the use and interpretation of certain test criteria for purposes of statistical inference. *Biometrika* 1928; **20A**, Part I: 175–240, Part II: 263–294.
16. Hindmarsh PC, Cole TJ. Height monitoring as a diagnostic test. *Archives of Disease in Childhood* 2004; **89**(4): 296–297.
17. Spiegelhalter DJ, Knill-Jones RP. Statistical and knowledge-based approaches to clinical decision-support systems, with an application to gastroenterology (with discussion). *Journal of Royal Statistical Society*, *Series A* 1984; **147**:35–76.



DOI: 10.1002/sim