Published in final edited form as:

Stat Med. 2011 May 30; 30(12): 1455–1465. doi:10.1002/sim.4189.

A two-part model for reference curve estimation subject to a limit of detection

Z. Zhang^{a,*}, O. Y. Addo^b, J. H. Himes^b, M. L. Hediger^a, P. S. Albert^a, A. L. Gollenberg^a, P. A. Lee^c, and G. M. Buck Louis^a

^aDivision of Epidemiology, Statistics and Prevention Research, Eunice Kennedy Shriver National Institute of Child Health and Human Development, Bethesda, Maryland, USA

^bDivision of Epidemiology and Community Health, University of Minnesota, Minneapolis, Minnesota, USA

^cDepartment of Pediatrics, Penn State College of Medicine, The Milton S. Hershey Medical Center, Hershey, Pennsylvania, USA

Abstract

Reference curves are commonly used to identify individuals with extreme values of clinically relevant variables or stages of progression which depend naturally on age or maturation. Estimation of reference curves can be complicated by a technical limit of detection (LOD) that censors the measurement from the left, as is the case in our study of reproductive hormone levels in boys around the time of the onset of puberty. We discuss issues with common approaches to the LOD problem in the context of our pubertal hormone study, and propose a two-part model that addresses those issues. One part of the proposed model specifies the probability of a measurement exceeding the LOD as a function of age. The other part of the model specifies the conditional distribution of a measurement given that it exceeds the LOD, again as a function of age. Information from the two parts can be combined to estimate the identifiable portion (i.e., above the LOD) of a reference curve and to calculate the relative standing of a given measurement above the LOD. Unlike some common approaches to LOD problems, the two-part model is free of untestable assumptions involving unobservable quantities, flexible for modeling the observable data, and easy to implement with existing software. The method is illustrated with hormone data from the Third National Health and Nutrition Examination Survey.

Keywords

centile curve; identifiability; LMS; missing data; quantile regression; z-score

1. Introduction

Reference curves, also known as centile curves, are plots of selected percentiles of a clinically relevant variable (such as weight or height) over a range of values for an important covariate, often age. Such curves are routinely used in screening and clinical care to identify individuals with extreme values who may require further examination or follow-up. There is a variety of statistical methods for estimating reference curves, many of which are reviewed by Wright and Royston [1,2]. This literature includes a class of nonparametric methods that

permit an arbitrary form of the distribution at each specific age [3,4,5,6,7,8,9,10]. These methods are robust against possible misspecification of a parametric model. On the other hand, they often require a large sample size, especially for estimating extreme percentiles. Furthermore, they do not provide a closed-form expression for the relative standing of an arbitrary measurement, which is usually desired in clinical practice. Closed-form formulas are usually based on parametric models for the distribution at each specific age, although the dependence of model parameters on age may be parametric or nonparametric [11,12,13,14,15,16,17]. While convenient to use, the parametric methods require careful modeling of the response distribution, for which a variety of modeling techniques such as transformations have been developed [1,2]. In particular, the LMS method of Cole and Green [14] offers remarkable flexibility by allowing the Box-Cox transformation parameter (L), the median (M) and the approximate coefficient of variation (S) to vary with age in a smooth but otherwise arbitrary manner.

In this paper, we consider how to deal with a limit of detection (LOD) that censors the measurement from the left. This problem has complicated our effort to construct reference curves for reproductive hormone levels in prepubertal and early pubertal boys. The subjects in this investigation were participants of the Third National Health and Nutrition Examination Survey (NHANES III, 1988–1994). The survey included 1767 boys aged 6–11.99 years, of which 839 had available stored residual sera samples. The available sera samples were subsequently assayed for luteinizing hormone, testosterone and inhibin B using immunoassay techniques. As one might expect given their ages, a fair number of these measurements were below the corresponding LODs, especially for younger boys. For example, the measured testosterone level is below the LOD (0.035 ng/ml) for about 30% of boys between 6 and 7 years of age. Appropriate analysis of the NHANES III hormone data is complicated by various issues with the existing methods.

It is comforting to know that substituting the LOD for a missing value below the LOD, as is commonly done, changes the distribution but not the percentiles above the LOD. This implies that a nonparametric analysis is valid if we focus on portions of reference curves above the LOD. However, these methods usually do not provide closed-form formulas, with the exception of the CDC method [17] (so named because it was used to produce growth charts by the United States Centers for Disease Control and Prevention). The CDC method involves grouping subjects into narrow intervals of age, calculating empirical percentiles in each group, smoothing them across age groups, and then using the smoothed percentiles to estimate the LMS parameters in each age group. The CDC method is insensitive to the LOD problem, at least for percentiles above the LOD, but it requires that each age group be homogeneous enough for the results to be interpretable and at the same time include enough subjects for the estimates to be stable. For the NHANES III hormone data, it makes clinical sense to work with half-year intervals for age [18], and in each interval 100 subjects are usually needed as a rule of thumb [19]. This suggests that the CDC method may require a sample comprising at least 1200 boys with hormone data (in excess of the 839 available boys).

The parametric methods that do not require grouping do assume that the measurement is continuously distributed. These methods are not directly applicable to the NHANES III hormone data, which contain a non-negligible discrete component due to the LOD. This misspecification issue can be addressed by including a likelihood component that explicitly acknowledges that the missing values are below, not equal to, the LOD [20,21]. This maximum likelihood approach is commonly used to handle missing data and yields efficient inference when the model is correctly specified. In the present context, however, the approach is not readily available and the associated model checking techniques have yet to be developed.

In this paper, we propose a simple two-part model for reference curve estimation that addresses the LOD problem in a natural and transparent way. One part of the model concerns the probability of exceeding the LOD as a function of age. The other part of the model characterizes the distribution of the hormone level above the LOD, again as a function of age. Combining information from the two parts yields estimates of desired percentiles above the LOD. The two-part model effectively separates the two distinct aspects of the observed data: detectability (a dichotomous outcome) and distribution above the LOD (a continuous component). No modeling is required for the unobserved data (below the LOD), and consequently no inference will be made about the below-LOD percentiles (other than the fact that they are below the LOD). The proposed approach can be easily implemented with existing software. It can be regarded as an extension of the aforementioned maximum likelihood approach with added flexibility.

The remainder of the paper is organized as follows. In the next section, we set up the notation and discuss the LOD problem. In Section 3, we present the two-part model and describe the associated estimation procedure. The proposed approach is then illustrated with the NHANES III hormone data and compared with other methods in Section 4. The paper concludes with a discussion in Section 5.

2. Reference curves and the LOD

Let Y and T denote the variable of interest (e.g., hormone level) and the age of a subject, respectively. We assume Y is positive and continuous at each age, and write $F(y|t)$ for the conditional distribution function of Y given $T = t$. For $0 < p < 1$, the 100 p th reference curve is just $F^{-1}(p|t)$ considered as a function of t . An individual who falls below a low reference curve (such as the 2.5th) or above a high one (such as the 97.5th) will be considered unusual and may require further examination. Sometimes the relative standing of a given measurement $Y = y$ at $T = t$ is expressed as $z = \Phi^{-1}(F(y|t))$, where Φ denotes the standard normal distribution function. Obviously, the z -score conveys the same information as the value of the distribution function. In clinical medicine, there is much interest in extreme reference curves and closed-form formulas for the distribution function.

Reference curves can be estimated using parametric or nonparametric methods. Parametric models are convenient for estimating extreme percentiles and for deriving closed-form formulas. Here a model $F(\cdot|t; \theta)$ is considered parametric if the submodel $F(\cdot|t; \theta)$ is parametric at each t ; the dependence on t may be parametric as in some of Royston's models or nonparametric. For example, under the LMS model of Cole and Green [14], θ consists of smooth curves L , M and S such that, given $T = t$, a Box-Cox transformation of Y of the form

$$Z = \begin{cases} \frac{\{Y/M(t)\}^{L(t)} - 1}{L(t)S(t)}, & L(t) \neq 0 \\ \frac{\log\{Y/M(t)\}}{S(t)}, & L(t) = 0 \end{cases} \quad (1)$$

follows a standard normal distribution. Here $M(t)$ gives the median of Y and $S(t)$ is the approximate coefficient of variation of Y for $L(t) \approx 1$. The L , M and S curves can be estimated by maximizing a penalized likelihood [14]. Of course, the validity of the parametric approach depends on correct specification of the model. A misspecified model can obscure important features of the data and result in inconsistent estimates. This concern has motivated development of nonparametric methods. For example, the double kernel method of Li et al. [10] requires no distributional assumptions and only two smoothers: a kernel k_t for T with bandwidth h_t and a distribution function K_y with bandwidth h_y for kernel

smoothing in the y -axis. Given the data $\{(T_i, Y_i) : i = 1, \dots, n\}$ from n independent subjects, the double kernel estimate of $F(y|t)$ is given by $\sum_{i=1}^n w_i^L(t) K_y(h_y^{-1}(y - Y_i))$, where

$$w_i^L(x) = w_i^k \left\{ 1 + \frac{(x - \bar{x}^k)(x_i - \bar{x}^k)}{\sum_{j=1}^n w_j^k(x)(x_j - \bar{x}^k)^2} \right\}$$

with $\bar{x}^k = \sum_{i=1}^n w_i^k(x) x_i$ and $w_j^k(x) = k_t(h_t^{-1}(x - x_j)) / \sum_{l=1}^n k_t(h_t^{-1}(x - x_l))$. Reference curves can be obtained by inverting the estimated distribution function. However, the estimated distribution function, though in closed form, is of limited utility to a clinician trying to determine the relative standing of a given individual.

Now suppose Y is observed only if it exceeds an LOD, say $d > 0$. A crude way to deal with the LOD issue is to simply replace each missing value of Y by d , essentially defining a pseudo-observation $Y^* = \max\{Y, d\}$. It is convenient that Y^* and Y have the same percentiles above d . Thus, if an appropriate method is applied to Y^* , the resulting reference curves will be valid for Y as well if we ignore the portions below d . This is the case for nonparametric methods that allow a mixed distribution (with both discrete and continuous components). However, there may be a misspecification issue with direct application of standard parametric models to the Y^* -values, because the models are for continuous data and Y^* has a discrete component. It is certainly possible to impute the missing value of Y in a continuous manner. However, the choice of an imputation model can be rather arbitrary, and its impact on the subsequent analysis may not be transparent.

A more sensible parametric approach is to specify a model for Y and maximize a likelihood that explicitly accounts for the missingness of some Y -values. This approach has been used to address LOD problems [20,21], and can be applied to the present setting as follows. Write $R = I(Y > d)$, where $I(\cdot)$ is the indicator function. Then the observed data based on n independent subjects can be thought of as $(T_i, R_i, R_i Y_i)$, $i = 1, \dots, n$, because Y_i can be recovered from $R_i Y_i$ if and only if $R_i = 1$. With $f(y|t; \theta)$ denoting the density of $F(y|t; \theta)$, the likelihood for θ is given by

$$\prod_{i=1}^n f(Y_i|T_i; \theta)^{R_i} F(d|T_i; \theta)^{1-R_i}, \quad (2)$$

which will need to be penalized if the LMS model is adopted. Maximizing this likelihood, or a penalized version of it, yields an estimate of θ that is consistent and efficient, assuming that the model $F(y|t; \theta)$ is correctly specified. This is a viable approach in principle, but there are several practical issues. Firstly, the method is not available in existing software and its implementation will require considerable effort, especially for a very flexible model such as the LMS model. Secondly, model checking techniques have yet to be developed, as some Y -values are not observed. This is a real issue because model checking is an essential part of reference curve estimation, where the estimates depend heavily on model specification (degrees of polynomials in Royston's models and smoothing parameters in the LMS model). Thirdly, some care must be taken when interpreting the results. Because the model $F(y|t; \theta)$ is specified for the entire distribution of Y , a careless analyst may be tempted to believe that the entire distribution has been estimated once an estimate of θ is substituted into the model. However, a part of the distribution (i.e., the tail distribution below the LOD) is not empirically identifiable, and that part of the model cannot be verified or refuted using the

observed data, even with perfect model checking techniques. Any inference about that part of the distribution is simply an extrapolation of the observed data in a direction dictated by the modeling assumptions. Mindful of this limitation, a thoughtful analyst might choose to ignore the distribution of Y given $Y \leq d$ and only interpret the rest of the model pertaining to the observed data. Our proposal in the next section extends this pragmatic approach in a way that also addresses the first two issues.

3. The two-part model

We now propose an approach that retains the advantages of parametric modeling and avoids the difficulties with the approach based on (2). The key to the proposed approach is to restrict the modeling effort to the observed data alone. For a generic subject, the observable can be written as (T, R, RY) , and the relevant part of the distribution can be factorized as

$$[R, RY|T] = [R|T][RY|T, R], \quad (3)$$

where $[\cdot|\cdot]$ denotes a generic conditional distribution. The last term in equation (3) is trivial if $R = 0$ (i.e., $Y \leq d$), and we are only interested in modeling the two non-trivial parts: $[R|T]$ and $[RY|T, R = 1] = [Y|T, Y > d]$. The overall model is therefore referred to as a two-part model, and it differs from the model $F(y|t; \theta)$ underlying (2) in that both parts of our model can be directly identified from the observed data. Similar models have been used to handle semicontinuous data in various contexts, such as medical cost data (modeled as a mixture of zero and a log-normal distribution) [22,23,24] and emesis volume data in an acupuncture clinical trial [25]. As far as we know, this two-part model approach has not been applied previously to reference curve estimation.

One part of our model, the dichotomous part, is given by

$$P(R=1|T=t) = P(Y>d|T=t) = \pi(t; \beta), \quad (4)$$

where π is a known function and β is an unknown finite-dimensional parameter. A common choice for (4) would be a logistic regression model where the dependence on age could be parametric (through linear terms) or nonparametric (as in a generalized additive model). The other part of the model, the continuous part concerning the Y -values above the LOD, can be written as

$$P(Y \leq y|T=t, Y>d) = G(y|t; \gamma). \quad (5)$$

Since we are modeling continuous data, we can take $G(y|t; \gamma)$ to be any one of the commonly used parametric models mentioned earlier. The new notation here, $G(y|t; \gamma)$ as opposed to $F(y|t; \theta)$, emphasizes the fact that we are now modeling a different aspect of the data. Information from the two parts, (4) and (5), can be combined through the following identity:

$$P(Y \leq y|T=t) = P(Y \leq d|T=t) + P(Y>d|T=t)P(Y \leq y|T=t, Y>d), \quad y>d.$$

In terms of parameters in the two-part model, the relative standing of a given measurement $Y = y$ at $T = t$ can be expressed as

$$F(y|t) \begin{cases} = 1 - \pi(t;\beta) + \pi(t;\beta)G(y|t;\gamma), & y > d; \\ \leq 1 - \pi(t;\beta), & y \leq d. \end{cases} \quad (6)$$

If a z-score is desired, it can be obtained by applying Φ^{-1} to the above quantity. Note that $F(y|t)$ is not fully specified for $y \leq d$. Inverting the distribution function yields the age-specific percentiles:

$$F^{-1}(p|t) \begin{cases} = G^{-1}\left\{\frac{p + \pi(t;\beta) - 1}{\pi(t;\beta)} \middle| t; \gamma\right\}, & p > 1 - \pi(t;\beta); \\ \leq d, & p \leq 1 - \pi(t;\beta). \end{cases} \quad (7)$$

The inequalities in the preceding displays should be understood as a truthful representation of the available data rather than a drawback of the model. Indeed, percentiles below the LOD and relative standings of $y \leq d$ are not completely identified from the data alone. The fact that these quantities can be identified and estimated under the model $F(y|t; \theta)$ merely illustrates the strength of the modeling assumptions [26,27]. Even when the model $F(y|t; \theta)$ is strongly supported by external information, an analysis based on the two-part model can help to clarify the contributions of the different sources of information (internal versus external).

This two-part modeling approach overcomes the difficulties associated with the approach based on (2) without losing its intuitive appeal. In fact, our proposal can be regarded as an extension of the latter approach with added flexibility in the following sense. If we specify a model $F(y|t; \theta)$ and focus on the part of the model relevant to the observed data, as indicated earlier, then we are effectively working with a two-part model given by

$$\begin{aligned} P(R=1|T=t) &= 1 - F(d|t; \theta) =: \pi(t; \theta), \\ P(Y \leq y|T=t, Y > d) &= \frac{F(y|t; \theta) - F(d|t; \theta)}{1 - F(d|t; \theta)} =: G(y|t; \theta), \end{aligned}$$

corresponding to (4) and (5), respectively. Thus the relevant part of the model $F(y|t; \theta)$ is really a constrained two-part model where both parts are governed by the same set of parameters (θ). The likelihood (2) can now be rewritten as

$$\prod_{i=1}^n \left\{ \frac{f(Y_i|T_i; \theta)}{1 - F(d|T_i; \theta)} \right\}^{R_i} \{1 - F(d|T_i; \theta)\}^{R_i} F(d|T_i; \theta)^{1-R_i} = \left\{ \prod_{i: R_i=1}^n g(Y_i|T_i; \theta) \right\} \times \prod_{i=1}^n \pi(T_i; \theta)^{R_i} \{1 - \pi(T_i; \theta)\}^{1-R_i},$$

where $g(y|t; \theta)$ is the density of $G(y|t; \theta)$. The right-hand side is exactly the joint likelihood for the two-part model. Its maximization is complicated by the constraint due to a common set of parameters. The proposed approach removes the constraint with separate parameters for the two parts, thereby simplifying the maximization of the likelihood. Without the constraint, the proposed model should be more flexible and potentially more capable of revealing important features of the data. On the other hand, there may be a loss of efficiency because constrained inference is generally more efficient when the constraint actually holds. One would need to be very confident about the constraint, though, in order to take advantage of the efficiency gain.

It is worth noting that a simple transformation (subtraction by the LOD) may help in fitting model (5). This part concerns the subset of data, $\{Y_i : Y_i > d\}$, which are bounded below by d . Some candidate models, such as the LMS model, may assume a range of all positive numbers. Direct fitting of such models may lead to unsatisfactory goodness-of-fit, and the estimated distribution, $G(\cdot|t; \hat{\gamma})$, may extend below d , which would be inconvenient. These problems can be avoided by working with the transformed data, $\{Y_i - d : Y_i > d\}$. After fitting the transformed data to an appropriate model, we will need to perform a “back transformation” for the subsequent calculations. This will be illustrated with the NHANES III hormone data. Aside from this technicality, both parts of the model, (4) and (5), can be specified using standard modeling techniques and estimated using existing procedures. The resulting estimates can then be substituted into expressions (6) and (7). The bootstrap methodology can be used to obtain standard errors, confidence intervals and confidence bands, if desired.

4. A case study

The NHANES III survey was designed to obtain nationally representative information on the health and nutritional status for the U.S. population. The data were collected under a complex, multistage, probability sampling design with oversampling of certain subpopulations. Sampling weights are available to account for oversampling and survey nonresponse and allow generalization to the target population. As mentioned earlier, the limited availability of hormone information restricts our analysis to a subset of the NHANES III sample. Of the 1767 boys aged 6–12 years, 839 had stored residual sera samples, and 825 contributed information on testosterone (the focus of this illustration). To account for this extra selection mechanism, we fitted a logistic regression model for predicting the availability of the desired hormone information and calculated inverse probability weights [28,29]. The model was based on race/ethnicity and general health status, which were found to be the most important predictors in a variable selection process. Assuming the hormone information is missing at random, which seems reasonable given the large amount of covariate information available, our analysis can be generalized to the target population by multiplying the inverse probability weights with the original sampling weights.

Our two-part model for testosterone was implemented using the modified weights for each specific analysis. The dichotomous part, given by equation (4), was specified as a logistic regression model with age as the only linear term. This model was preferred over a generalized additive model because it respects a natural monotonicity (testosterone level is expected to increase with age and so is the probability of exceeding the LOD). The model was fitted in a weighted logistic regression analysis using the modified weights described earlier. Figure 1 compares the predicted probability of exceeding the LOD as a function of age with observed proportions in age bins (weighted by the modified weights). The model appears to fit the data reasonably well. Little improvement in the goodness-of-fit can be achieved by adding another term, such as log-transformed, squared or square root of age. The fitted model can be expressed as

$$\hat{\pi}(t) = \{1 + \exp(1.829 - 0.035t)\}^{-1},$$

where t denotes age in months. For the continuous part, given by equation (5), an LMS model was fitted to the transformed data, $\{Y_i - d : Y_i > d\}$, where $d = 0.035$ ng/ml is the LOD for testosterone. This analysis was also performed using the modified weights. The equivalent degrees of freedom (e.d.f.: 3 for L ; 4 for M ; 3 for S) were chosen with the help of Q-tests and worm plots [30, 31, 32].

Combining information from the two parts, the relative standing of a given measurement $Y = y$ at age t (months) is estimated as

$$\widehat{F}(y|t) = \begin{cases} 1 - \widehat{\pi}(t) + \widehat{\pi}(t) \Phi \left[\frac{\{(y-d)/\widehat{M}(t)\}^{\widehat{L}(t)} - 1}{\widehat{L}(t)\widehat{S}(t)} \right], & \widehat{L}(t) \neq 0; \\ 1 - \widehat{\pi}(t) + \widehat{\pi}(t) \Phi \left[\frac{\log\{(y-d)/\widehat{M}(t)\}}{\widehat{S}(t)} \right], & \widehat{L}(t) = 0, \end{cases}$$

for $y > d$. Here $\widehat{L}(t)$, $\widehat{M}(t)$ and $\widehat{S}(t)$ denote estimates from the LMS analysis. If the measurement is censored by the LOD, then $F(y|t)$ is estimated to be below $1 - \widehat{\pi}(t)$. If z -scores are needed, they can be obtained by applying Φ^{-1} to the above quantities. By inverting the estimated distribution function, the 100 p th percentile of testosterone level at age t is estimated to be below d if $p \leq 1 - \widehat{\pi}(t)$; otherwise we have the estimate

$$\widehat{F}^{-1}(p|t) = \begin{cases} d + \widehat{M}(t) \{1 + \widehat{L}(t)\widehat{S}(t)z_q\}^{1/\widehat{L}(t)}, & \widehat{L}(t) \neq 0; \\ d + \widehat{M}(t) \exp\{\widehat{S}(t)z_q\}, & \widehat{L}(t) = 0, \end{cases}$$

where $q = \{p + \widehat{\pi}(t) - 1\}/\widehat{\pi}(t)$, and z_q is the 100 q th percentile of the standard normal distribution.

Estimates of selected reference curves based on the two-part model are compared with those based on a standard LMS analysis and a nonparametric double kernel method [10], both graphically and numerically in terms of goodness-of-fit. Here goodness-of-fit is assessed by comparing the proportion of data points falling below a curve with the nominal level of the curve. The standard LMS analysis and the double kernel method are implemented with all values below the LOD replaced by the LOD. The standard LMS analysis uses the same e.d.f.'s as in the continuous part of the two-part estimation. The double kernel method is described briefly in Section 2 and is implemented with a median correction. The double kernel method involves kernel smoothing in both directions (age and testosterone level) and thus requires two smoothers. For this example we use a Gaussian kernel for age and a uniform kernel for testosterone level as in Li et al. [10], and consider a range of values for the bandwidth parameters. In our experiments, the bandwidth for age takes the form $h_t = 2^{-b_t} \text{sd}(T)$, where sd denotes sample standard deviation and b_t ranges from -3 to 6 . Similarly, the bandwidth for testosterone level is given by $h_y = 2^{-b_y} \text{sd}(Y^*)$, where Y^* is a truncation of Y at the LOD introduced in Section 2 and b_y ranges from 0 to 6 . Selection of (b_t, b_y) is facilitated by the empirical finding that the appearance of the curve depends mostly on b_t while the goodness-of-fit is largely a function of b_y alone, as can be seen in Table 1 and Figure 2.

Table 1 compares the proposed method with the other methods in terms of goodness-of-fit as defined in the last paragraph. The comparison is done for five percentiles (10th, 25th, 50th, 75th, 90th) as well as the mean squared error (MSE), which is the average (across the five percentiles) of the squared difference between the actual proportion and the nominal level. Each method in the table involves smoothing and is specified through a set of smoothing parameters: e.d.f.'s for the standard LMS analysis and the continuous part of the two-part model, and (b_t, b_y) for the double kernel method. For each method we have experimented with different values of the smoothing parameters, perhaps more thoroughly for the double kernel method, which can be iterated automatically. Nonetheless, Table 1 does include results from LMS analyses with more or less smoothing than the actual analysis (e.d.f. = 3 for L ; 4 for M ; 3 for S), by decreasing or increasing the e.d.f.'s by 1 in each component. For the two-part model, the original set of e.d.f.'s chosen through Q-tests

and worm plots is associated with the smallest MSE among the 3 sets considered, although the differences are very small both in terms of MSE and for individual percentiles. For the standard LMS method, the differences between the different e.d.f.'s are also very small. Comparing the two parametric methods, the two-part model appears to fit the data slightly better than the standard LMS method in the sense that the associated MSE is consistently smaller. This is not surprising because the two-part model has two more parameters (in the logistic regression model for the dichotomous part) if the same set of e.d.f.'s is used by the two methods. On the other hand, higher e.d.f.'s do not necessarily lead to a smaller MSE, so the improved goodness-of-fit for the two-part model may not be attainable by simply increasing the e.d.f.'s. Generally, the standard LMS model appears to fit better at higher percentiles (75th and 90th), while the two-part model tends to fit better at lower ones (25th and 50th). For the 10th percentile, although the results in the table do not favor either method, we should remember that, for the standard LMS method, the "proportion below" is calculated with the below-LOD values replaced by the LOD. This substitution works to bias the observed proportion downward, so the "true proportion" below the 10th percentile curve should be higher than observed for the standard LMS method. The two-part model is not subject to this bias because it only estimates portions of reference curves above the LOD. Thus it is highly likely that the two-part model actually fits better at the 10th percentile. After all, the two-part model is designed and constructed to deal with the LOD issue facing low percentiles. The lowest MSE's in Table 1 are attained by the double kernel method. This should be interpreted carefully as we are not making a systematic comparison of parametric methods and nonparametric methods in general; the focus here is on comparing the proposed method with the other methods. Nonetheless, it makes intuitive sense that a nonparametric method could be tuned to fit the data better than a parametric method. The goodness-of-fit for the double kernel method appears to depend more on smoothing in the y-axis (governed by b_y) than on smoothing over age (governed by b_t). This conclusion is based on all available results for the double kernel method, of which only a small subset is shown in Table 1 to illustrate the trends. Among all the scenarios we have explored, the smallest MSE is attained at $(b_t, b_y) = (-1, 3)$, although the "global" minimum MSE (0.049) is not much lower than the "local" minimum (0.053) for $b_t = -2$ (attained at $b_y = 3$).

Figure 2 shows reference curves estimated using the three methods. The curves shown for the two-part model are the same ones generated in the actual analysis using a set of e.d.f.'s selected through Q-tests and worm plots and corresponding to the lowest MSE in the upper portion of Table 1. The curves shown for the standard LMS method are based on the same set of e.d.f.'s. Not shown in Figure 2 are parametric curves based on the other two sets of e.d.f.'s considered in Table 1. Those curves differ from the ones shown very slightly, with a little more or less smoothness, of course, depending on the smoothing parameters. For the double kernel method, preliminary plots suggest that the appearance of the curves depends mostly on b_t , the smoothing parameter for age. So it seems reasonable to compare curves based on different values of b_t , each coupled with a value of b_y that minimizes the MSE locally. Shown in Figure 2 are two sets of such curves with $(b_t, b_y) = (-1, 3)$ and $(1, 2)$, the former corresponding to the global minimum MSE. These curves illustrate the general trend of increasing wiggleness with less smoothing (i.e., larger b_t). Not shown in Figure 2 are curves for $b_t = -2$, which are virtually indistinguishable from those for $b_t = -1$, and curves for $b_t = 0$, which are intermediate between the two sets shown. When all sets of curves in Figure 2 are compared with each other, there does appear to be a fair amount of disagreement. The differences are smaller for the median, perhaps with the exception of the wiggly one (double kernel with $(b_t, b_y) = (1, 2)$). For the higher percentiles, the results in Table 1 suggest that the standard LMS curves and the double kernel curves should be considered more credible. For the low percentiles, the two-part model fits better than the standard LMS model, and at least as well as the double kernel curves. Visual inspection of the curves for lower percentiles lends further support to the two-part model. For example, all

of the other methods produce 25th reference curves that are entirely above the LOD, which is counter-intuitive when the proportion of 6-year-old boys exceeding the LOD is estimated to be below 75% (see Figure 1). Similarly, the 10th reference curves from the other methods are at least partially above the LOD for boys 7 years and older, even though the estimated proportion of 7-year-old boys exceeding the LOD is far below 90%.

In summary, the two-part model appears effective for dealing with the LOD issue in estimating lower percentiles using the NHANES III data. The downside is a relative lack-of-fit for higher percentiles. Overall, the two-part model fits the data slightly better than the standard LMS model using the same e.d.f.'s, but not as well as a nonparametric method such as the double kernel method with carefully chosen smoothing parameters. Like the standard LMS model, the two-part model leads to closed-form formulas for $\hat{F}(y|t)$ that are convenient for clinical use and that are usually unavailable from nonparametric methods.

5. Discussion

LOD problems have become increasingly common, especially in studies involving biomarkers. The appropriate statistical approach to an LOD problem will certainly depend on the context and the research question. There are, however, some principles that we believe are widely applicable. We believe that, to the extent possible, modeling assumptions should accommodate important features of the data in a way that is transparent, flexible and empirically verifiable. The two-part model proposed here is just one step toward that direction. The proposed model is very simple, conceptually, mathematically and computationally. We regard this simplicity as an advantage over some alternative methods such as the maximum likelihood approach based on (2). While developed for reference curve estimation, the main idea of the two-part model may extend to other areas in medicine.

The LOD problem is really a special type of missing data. The missingness here is actually non-ignorable in the sense that it depends solely on the missing outcome, although the missing-data mechanism is simple and known. There are two general modeling strategies for missing data: selection models and pattern mixture models [33]. Selection models describe the distribution of quantities that could have been observed as well as the mechanism by which some quantities might be missing. The likelihood (2) can be seen as a selection model where $F(d|T; \theta)$ is the probability of Y_i being missing. (For $Y_i > d$, the probability of being observed is 1 and does not contribute to the likelihood.) The two-part model is related to, but different from, a pattern mixture model based on the factorization

$$[R, Y|T] = [R|T][Y|T, R].$$

The difference between this and equation (3) is that we are not interested in the distribution $[Y|T, R = 0] = [Y|T, Y \leq d]$ or the resulting mixture for $[Y|T]$. So the proposed approach is not really a pattern mixture approach. Usually, with missing data, some untestable assumptions must be made in order to identify the distribution of the original (possibly missing) variable. In the above pattern mixture model, the untestable part is the distribution $[Y|T, R = 0] = [Y|T, Y \leq d]$. Without imposing any untestable assumptions, the two-part model does not lead to complete identification of the distribution $[Y|T]$ and the associated mean and variance. However, for reference curve estimation, this is not a serious limitation, as much of the inferential target (i.e., percentiles above the LOD) remains identifiable and estimable under the two-part model. In studies like the NHANES III hormone example, there is hardly any interest in estimating percentiles below the LOD.

The analysis in Section 4 is aimed at generating reference curves for clinical use. Such curves are usually published without confidence bands (see, for example, the growth charts

at <http://cdc.gov/growthcharts/>). If a confidence interval or band is needed, say to compare a particular percentile with some benchmark, it can be obtained by bootstrapping the subjects in the sample. There are some caveats, though. Firstly, to make a bootstrap procedure feasible, the two-part estimation must be implemented in a computational environment that permits iterations, such as R. To our knowledge, an R package is not yet available for the LMS analysis or Royston's methods. Thus, some programming may be required to implement one of those methods in a suitable programming language. Secondly, the continuous part of the two-part model usually requires some exploratory analysis to choose the e.d.f.'s (for the LMS model) or specify suitable polynomials (for some of Royston's models). Such exploratory analysis will need to be automated in a bootstrap procedure. Lastly, for the NHANES III hormone data, it should be kept in mind that the inverse probability weights are not known in advance but estimated from the same data. To account for this source of variability, one would bootstrap all 1767 boys in the specified age range (with and without hormone data), and for each bootstrap sample recalculate the inverse probability weights and refit the two-part model.

Acknowledgments

This research was supported by the Intramural Research Program of the National Institutes of Health, *Eunice Kennedy Shriver* National Institute of Child Health and Human Development. The authors would like to thank the associate editor and two anonymous reviewers for insightful comments that have greatly improved the manuscript.

References

1. Wright EM, Royston P. A comparison of statistical methods for age-related reference intervals. *Journal of the Royal Statistical Society, Series A (Statistics in Society)*. 1997; 160:47–69.
2. Wright EM, Royston P. Calculating reference intervals for laboratory measurements. *Statistical Methods in Medical Research*. 1999; 8:93–112. [PubMed: 10501648]
3. Koenker R, Bassett G. Regression quantiles. *Econometrica*. 1978; 46:33–50.
4. Healy MJR, Rasbash J, Yang M. Distribution-free estimation of age-related centiles. *Annals of Human Biology*. 1988; 15:17–22. [PubMed: 3348588]
5. Himes J, Hoaglin DC. Resistant cross-age smoothing of age-specific percentiles for growth reference data. *American Journal of Human Biology*. 1989; 1:165–173.
6. Pan HQ, Goldstein H, Yang Q. Nonparametric estimation of age-related centiles over wide age ranges. *Annals of Human Biology*. 1990; 17:475–481. [PubMed: 2281939]
7. Wellek S, Merz E. Age-related reference ranges for growth parameters. *Methods of Information in Medicine*. 1995; 34:523–528. [PubMed: 8713769]
8. He X. Quantile curves without crossing. *The American Statistician*. 1997; 51:186–192.
9. Heagerty PJ, Pepe MS. Semiparametric estimation of regression quantiles with application to standardizing weight for height and age in US children. *Applied Statistics*. 1999; 48:533–551.
10. Li Y, Graubard BI, Korn EL. Application of nonparametric quantile regression to body mass index percentile curves from survey data. *Statistics in Medicine*. 2010; 29:558–572. [PubMed: 20013898]
11. Cole TJ. Fitting smoothed centile curves to reference data (with discussion). *Journal of the Royal Statistical Society, Series A (Statistics in Society)*. 1988; 151:385–418.
12. Thompson ML, Theron GB. Maximum likelihood estimation of reference centiles. *Statistics in Medicine*. 1990; 9:539–548. [PubMed: 2349405]
13. Royston P. Constructing time-specific reference ranges. *Statistics in Medicine*. 1991; 10:675–690. [PubMed: 2068420]
14. Cole TJ, Green PJ. Smoothing reference centile curves: the LMS method and penalized likelihood. *Statistics in Medicine*. 1992; 11:1305–1319. [PubMed: 1518992]
15. Altman DG. Construction of age-related reference centiles using absolute residuals. *Statistics in Medicine*. 1993; 12:917–924. [PubMed: 8337548]

16. Royston P, Wright EM. A method for estimating age-specific reference intervals ('normal ranges') based on fractional polynomials and exponential transformation. *Journal of the Royal Statistical Society, Series A (Statistics in Society)*. 1998; 161:79–101.
17. Kuczmarski RJ, Ogden CL, Guo SS, Grummer-Strawn LM, Flegal KM, Mei Z, Wei R, Curtin LR, Roche AF, Johnson CL. 2000 CDC growth charts for the United States: methods and development. *Vital Health Statistics, Series 11*. 2002; 246:1–190.
18. Himes, J.; Addo, OY.; Zhang, Z.; Gollenberg, AL.; Hediger, ML.; Lee, PA.; Louis, GMB. Reference curves for inhibin-B and testosterone for prepubertal children. Poster presentation at Pediatric Academic Societies' Annual Meeting; Vancouver, Canada. 2010.
19. Curtin, LR.; Chen, TC. Estimating extreme percentiles for BMI: minimum sample size required and sensitivity to kurtosis. Poster presentation at the Joint Statistical Meetings; Washington DC, USA. 2009.
20. Hughes JP. Mixed effects models with censored data with application to HIV RNA data. *Biometrics*. 1999; 55:625–629. [PubMed: 11318225]
21. Taylor DJ, Kupper LL, Rappaport SM, Lyles RH. A mixture model for occupational exposure mean testing with a limit of detection. *Biometrics*. 2001; 57:681–688. [PubMed: 11550915]
22. Duan N, Manning WG, Morris CN, Newhouse JP. A comparison of alternative models for the demand for medical care. *Journal of Business and Economic Statistics*. 1983; 1:115–126.
23. Zhou X-H, Tu W. Comparison of several independent population means when their samples contain log-normal and possibly zero observations. *Biometrics*. 1999; 55:645–651. [PubMed: 11318228]
24. Tian L, Huang J. A two-part model for censored medical cost data. *Statistics in Medicine*. 2007; 26:4273–4292. [PubMed: 17330248]
25. Albert PS, Shen J. Modelling longitudinal semicontinuous emesis volume data with serial correlation in an acupuncture clinical trial. *Applied Statistics*. 2005; 54:707–720.
26. Manski, CF. *Partial Identification of Probability Distributions*. New York: Springer-Verlag; 2003.
27. Zhang Z. Likelihood-based confidence sets for partially identified parameters. *Journal of Statistical Planning and Inference*. 2009; 139:696–710.
28. Horvitz DG, Thompson DJ. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*. 1952; 47:663–685.
29. Robins JM, Rotnitzky A, Zhao LP. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*. 1994; 89:846–866.
30. Royston P, Wright EM. Goodness-of-fit statistics for age-specific reference intervals. *Statistics in Medicine*. 2000; 19:2943–2962. [PubMed: 11042625]
31. van Buuren S, Fredriks M. Worm plot: a simple diagnostic device for modelling growth reference curves. *Statistics in Medicine*. 2001; 20:1259–1277. [PubMed: 11304741]
32. Pan HQ, Cole TJ. A comparison of goodness of fit tests for age-related reference ranges. *Statistics in Medicine*. 2004; 23:1749–1765. [PubMed: 15160406]
33. Little, RJA.; Rubin, DB. *Statistical Analysis with Missing Data*. 2nd ed. New York: Wiley; 2002.

Figure 1. Testosterone above the LOD: observed weighted proportions in age bins (bar chart) versus estimated probability as a function of age (dashed line).

Figure 2.

References curves for testosterone in prepubertal and early pubertal boys, constructed using the proposed two-part model, the standard LMS method and the double kernel method. The legend shows values of smoothing parameters: e.d.f.'s for the two-part model and the standard LMS model, and (b_t, b_y) for the double kernel method (see Section 4 for details).

Empirical comparison of reference curves constructed using the two-part model, the standard LMS method and the double kernel method, in terms of the proportion of data points falling below a curve and the mean squared error (MSE), the average squared difference between the actual proportion and the nominal level for several percentiles. Each method involves a set of smoothing parameters: e.d.f.'s for the two-part model and the standard LMS model, and (b_t, b_y) for the double kernel method (see Section 4 for details).

Table 1

Method	Smoothing Parameters	Proportion Below				MSE	
		10th	25th	50th	75th		90th
Two-part model							
	e.d.f. (L, M, S)						
	(2, 3, 2)	0.09	0.23	0.48	0.70	0.87	0.067
	(3, 4, 3)	0.09	0.23	0.48	0.70	0.87	0.066
	(4, 5, 4)	0.09	0.23	0.48	0.70	0.86	0.067
Standard LMS							
	e.d.f. (L, M, S)						
	(2, 3, 2)	0.11	0.20	0.46	0.73	0.89	0.069
	(3, 4, 3)	0.11	0.20	0.46	0.73	0.89	0.070
	(4, 5, 4)	0.11	0.21	0.45	0.73	0.88	0.072
Double kernel							
	(b_t, b_y)						
	(-2, 1)	0.03	0.18	0.52	0.77	0.89	0.108
	(-2, 2)	0.06	0.22	0.50	0.74	0.87	0.057
	(-2, 3)	0.07	0.23	0.49	0.72	0.87	0.053
	(-2, 4)	0.09	0.23	0.48	0.72	0.87	0.059
	(-1, 1)	0.03	0.18	0.52	0.77	0.89	0.108
	(-1, 2)	0.06	0.22	0.49	0.74	0.88	0.056
	(-1, 3)	0.07	0.24	0.49	0.73	0.87	0.049
	(-1, 4)	0.09	0.23	0.48	0.72	0.87	0.055
	(1, 1)	0.02	0.17	0.52	0.77	0.89	0.117
	(1, 2)	0.07	0.23	0.48	0.74	0.87	0.056
	(1, 3)	0.08	0.24	0.47	0.72	0.87	0.059
	(1, 4)	0.08	0.23	0.48	0.71	0.87	0.063