The use of multiple imputation for the accurate measurements of individual feed intake by electronic feeders

S. Jiao,*[1] F. Tiezzi,* Y. Huang,† K. A. Gray,† and C. Maltecca*

*Department of Animal Science, North Carolina State University, Raleigh 27695;
and †Smithfield Premium Genetics, Rose Hill, NC 28458

Downloaded from https://academic.oup.com/jas/article/94/2/824/4701700 by Università degli Studi di Firenze user on 24 February 2022

ABSTRACT: Obtaining accurate individual feed intake records is the key first step in achieving genetic progress toward more efficient nutrient utilization in pigs. Feed intake records collected by electronic feeding systems contain errors (erroneous and abnormal values exceeding certain cutoff criteria), which are due to feeder malfunction or animal–feeder interaction. In this study, we examined the use of a novel data-editing strategy involving multiple imputation to minimize the impact of errors and missing values on the quality of feed intake data collected by an electronic feeding system. Accuracy of feed intake data adjustment obtained from the conventional linear mixed model (LMM) approach was compared with 2 alternative implementations of multiple imputation by chained equation, denoted as MI (multiple imputation) and MICE (multiple imputation by chained equation). The 3 methods were compared under 3 scenarios, where 5, 10, and 20% feed intake error rates were simulated. Each of the scenarios was replicated 5 times. Accuracy of the alternative error adjustment was measured as the correlation between the true daily feed intake (DFI; daily feed intake in the testing period) or true ADFI (the mean DFI across testing period) and the adjusted DFI or adjusted ADFI. In the editing process, error cutoff criteria are used to define if a feed intake visit contains errors. To investigate the possibility that the error cutoff criteria may affect any of the 3 methods, the simulation was repeated with 2 alternative error cutoff values. Multiple imputation methods outperformed the LMM approach in all scenarios with mean accuracies of 96.7, 93.5, and 90.2% obtained with MI and 96.8, 94.4, and 90.1% obtained with MICE compared with 91.0, 82.6, and 68.7% using LMM for DFI. Similar results were obtained for ADFI. Furthermore, multiple imputation methods consistently performed better than LMM regardless of the cutoff criteria applied to define errors. In conclusion, multiple imputation is proposed as a more accurate and flexible method for error adjustments in feed intake data collected by electronic feeders.

Key words: electronic feeders, feed intake, multiple imputation

© 2016 American Society of Animal Science. All rights reserved. J. Anim. Sci. 2016.94:824–832
doi:10.2527/jas2015-9667

INTRODUCTION

Feed efficiency is a trait of primary economic importance in the swine industry. To achieve genetic progress toward more efficient nutrient utilization in pigs, obtaining accurate individual feed intake is essential. Computerized electronic feeding systems developed to automatically measure feed intake have greatly facilitated the data collection process. However, it has been known that feed intake data collected by those systems contain errors and outliers due to feeder malfunction and animal–feeder interaction (de Haer et al., 1992), which have been categorized by Eissen et al. (1998) and Casey et al. (2005). Due to the fact that simply discarding feed intake visits containing errors underestimates the true daily feed intake (**DFI**), a linear mixed model (**LMM**) has been proposed by Casey (2003) to adjust records containing errors after removing visits with missing values. However, applying LMM to a data set with extreme values remains challenging because those values tend to severely bias the estimates (Osborne and Overbay, 2004).

Multiple imputation was introduced by Rubin (1976) as a method with the very general task of predicting missing values. This approach has gained increasing popularity and has been implemented

[1]Corresponding author: sjiao@ncsu.edu
Received August 11, 2015.
Accepted November 30, 2015.

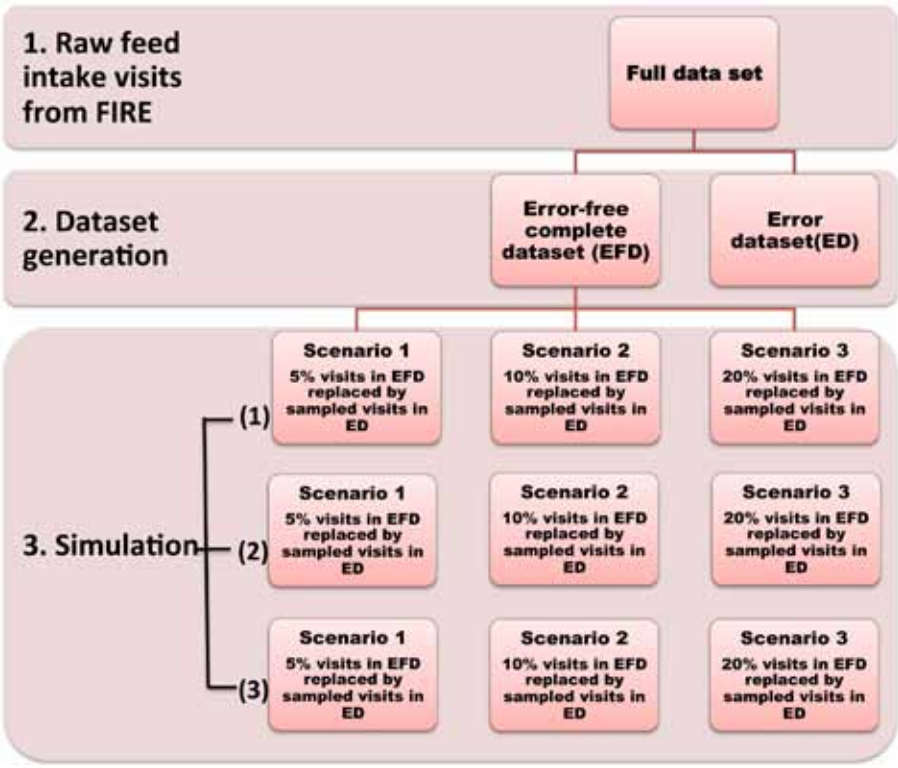

Figure 1. Working flow chart for data generation and simulation. The full data set contained 4,958,077 raw feed intake visits (without any editing or adjustment) collected using Feed Intake Recording Equipment (FIRE; Osborne Industries, Inc., Osborne, KS) from the year 2004 to 2013 for 14,901 pigs. The error-free complete data set (EFD) was generated from the full data set with animals meeting the requirements to enter, which were treated as "true" visits. The requirements for pigs to be selected to enter the EFD were 1) there were no error visits for the animal, 2) the animal had at least 2 feed intake visits to the feeder a day, 3) the testing period for each individual lasted at least 60 d, and 4) each contemporary group (concatenation of birth year, season, and house) had at least 15 pigs. After filtering, there were 17,908 feeding visits for the 100 selected pigs belonging to 4 contemporary groups. The data in the EFD were treated as "true" feed intake records. An error data set (ED) was generated by identifying errors in the full data set. The simulation was repeated 3 times (simulations (1), (2), and (3) in the figure) with different error cutoff criteria (the threshold values defining errors). Within each replication of simulation, 3 scenarios were simulated with 5, 10, and 20% visits in the EFD substituted by randomly sampled error visits from the ED.

in many areas of statistical analysis (Rubin, 1996; Allison, 2002). The key concept of this technique is the use of the distribution of the observed data to estimate a set of plausible values for the missing data.

Recent advances in software development provide opportunities to use MI (multiple imputation) in generating more accurate feed intake data collected by electronic systems.

The objective of this study was to evaluate the performance of 2 alternative implementations of multiple imputation, denoted as MI (multiple imputation) and MICE (multiple imputation by chained equation), in replacing errors and missing observations occurring in feed intake data, compared with the well-established LMM approach, under different simulated scenarios.

MATERIALS AND METHODS

Animal Selection and Data Set Generation

To mimic realistic error patterns, an error-free data set and the corresponding simulated data sets with different percentages of error visits were generated from real feed intake records collected by electronic feeders. The flow chart for the data creation and simulation process is summarized in Fig. 1.

Feed intake records used were collected from 2004 through 2013 using the Feed Intake Recording Equipment (**FIRE**; Osborne Industries, Inc., Osborne, KS) system in a Duroc nucleus herd owned by Smithfield Premium Genetics (Rose Hill, NC). Data included 4,958,077 feed intake visits for 14,901 animals with a testing period lasting, on average, 45 d. This data set containing raw feeding visits was defined as the "full data set" (**FD**). The detailed testing procedure to collect individual feed intake using FIRE can be found in Chen et al. (2010) and Jiao et al. (2014). Briefly, during the testing period, an average of 12 pigs housed together in a pen had 24-h access to feed with a single-spaced electronic feeder. When a visit to the feeder occurred, the pig identification number, date, entry feed weight (feed weight when a pig entered the feeder), exit feed weight, entry time, exit time, and pig BW were recorded. Quantities measured by the feeding system can be summarized into feed intake per visit (**FIV**; g), occupation time per visit (**OTV**; s), and feed intake rate per visit (**FRV**; g/min). Feed intake per visit

Downloaded from https://academic.oup.com/jas/article/94/2/824/4701700 by Università degli Studi di Firenze user on 24 February 2022

Table 1. Types of errors in feed intake visits from Feed Intake Recording Equipment (Osborne Industries, Inc., Osborne, KS) and rate of error for each error type in the full data set from the year 2004 to 2013

Error index	Error type[1]	Error definition[2]	Error rate (100%)[3]
1	FIV-low	FIV < –20 g for all visits	3.07
2	FIV-high	FIV > 2,000 g for all visits	4.46
3	FIV-0	FIV > 20 g or FIV < –20 g for visits with OTV = 0 s	0.00
4	OTV-low	OTV < 0 s for all visits	0.53
5	OTV-high	OTV > 3,600 s for all visits	0.10
6	FRV-high–FIV-low	FRV > 500 g/min for visits with 0 < FIV < 50 g	0.00
7	FRV-high	FRV > 350 g/min for visits with FIV > 50 g	2.20
8	FRV-0	FRV = 0 g/min for visits with OTV > 500 s	0.80

[1]Eight error types were proposed by Casey et al. (2005) and Eissen et al. (1998): FIV = feed intake per visit (g); OTV = occupation time per visit (s); FRV = feed intake rate per visit (g/min).

[2]The cutoff criterion were based on Casey et al. (2005) for different error types, which were chosen based on the feature of the feeder (error type 1 and 3), the biology of pig for feed intake (error type 4, 5, and 8), or the distribution of the variables FIV, OTV, or FRV (error type 2, 5, and 7).

[3]The error rate is computed for the full data set, where the overall error rate (the number of visits with at least 1 error/total visits) is 9.28%.

was computed as the difference of entry and exit feed weight. Similarly, the OTV was calculated as the difference between exit and entry time of the visit. Feeding rate per visit was defined as the ratio between FIV and OTV. For each of the parameters previously outlined, errors contained in each visit were defined as values more extreme than a predetermined cutoff value. Cutoff values frequently used by the industry for FIV, OTV, and FRV are those recommended by Casey (2003), whereas other cutoff values are generally based on knowledge of the feeders or biology of the pigs (Eissen et al., 1998; Casey, 2003). The 8 most commonly used error types occurring in the 3 variables FIV, OTV, and FRV are those based on studies conducted by Eissen et al. (1998) and Casey (2003). The error rates for these error types in the raw feed intake data set of the present study are displayed in Table 1. Two of these 8 types of feed-intake measurement errors were not detected in the current FD.

An error-free complete data set (**EFD**) and an error data set (**ED**) were generated by identifying missing feed intake values and each of the error types in each feeding visit event in the FD. The EFD was created as reference data set in simulation with all feeding visits considered accurately measured, whereas the ED contained only visits with errors or missing values. To reduce computation costs while preserving the general validity of the results, the EFD was generated as a subset of FD selecting 100 animals born in year 2013. Animals were selected to enter the EFD if 1) there were no error visits for the animal, 2) the animal had at least 2 feed intake visits to the feeder a day, 3) the testing period for each individual lasted at least 60 d, and 4) each contemporary group (concatenation of birth year, season, and house) had at least 15 pigs. After filtering, there were 17,908 feeding visits for the 100 selected pigs belonging to 4 contemporary groups. The data in the EFD were treated as true feed intake records without errors.

To mimic realistic error-occurring patterns, error visits were introduced into the EFD by masking the true values. Error events were randomly assigned, including one or several combination of errors in 1 visit. To achieve this goal, the simulated data were generated by randomly selecting true visits in the EFD and then substituting them with random samples of error visits from the ED. To assess the influence of different error rates (ratio of number of visits with errors over total visits) on error adjustment accuracy, 3 simulation scenarios were considered with error rates of 5, 10, and 20%, respectively (Fig. 1). For each scenario, 5 replicates were independently generated to eliminate any sparse randomness of adjustment results. As a result, there were 15 simulated data sets under the 3 simulated scenarios. The error rate for each error type for the 15 simulated data sets is presented in Table 2.

The cutoff criteria determining whether a visit contained an error were based on hands-on knowledge of the feeders, the criteria developed early (Eissen et al., 1998), or the distribution of variables such as FIV, OTV, and FRV (Casey, 2003). To verify that the choice of cutoff points had no impact on the effectiveness of the error adjustment methods evaluated, the outlined simulation was replicated 2 additional times, under more and less stringent cutoff criteria for FIV, OTV, and FRV. This was done by either doubling or halving the original error thresholds based on their empirical distribution (Supplemental Figure S1; see the online version of the article at http://journalofanimalscience.org). The corresponding error rates are shown in Fig. 2.

Statistical Analysis for Error Adjustment

Linear Mixed Model. The justification for adjusting error visits stems from the fact that simply discarding these visits would severely underestimate DFI (Casey, 2003), which was computed by summing FIV by testing day for each pig. A LMM to adjust error records in feed intake collected by FIRE was developed by Casey (2003) and was considered the conventional approach. Briefly, in this approach, percentage of errors, DFI, and daily occupation time summarized for visits of a certain error type were regressed on error-free DFI ($\mathbf{DFI_{ef}}$) as covariates to compute an adjusted value. In our data, a matrix representation of the model is as follows:

$$\mathbf{y = Xb + Zu + e,}$$

Downloaded from https://academic.oup.com/jas/article/94/2/824/4701700 by Università degli Studi di Firenze user on 24 February 2022

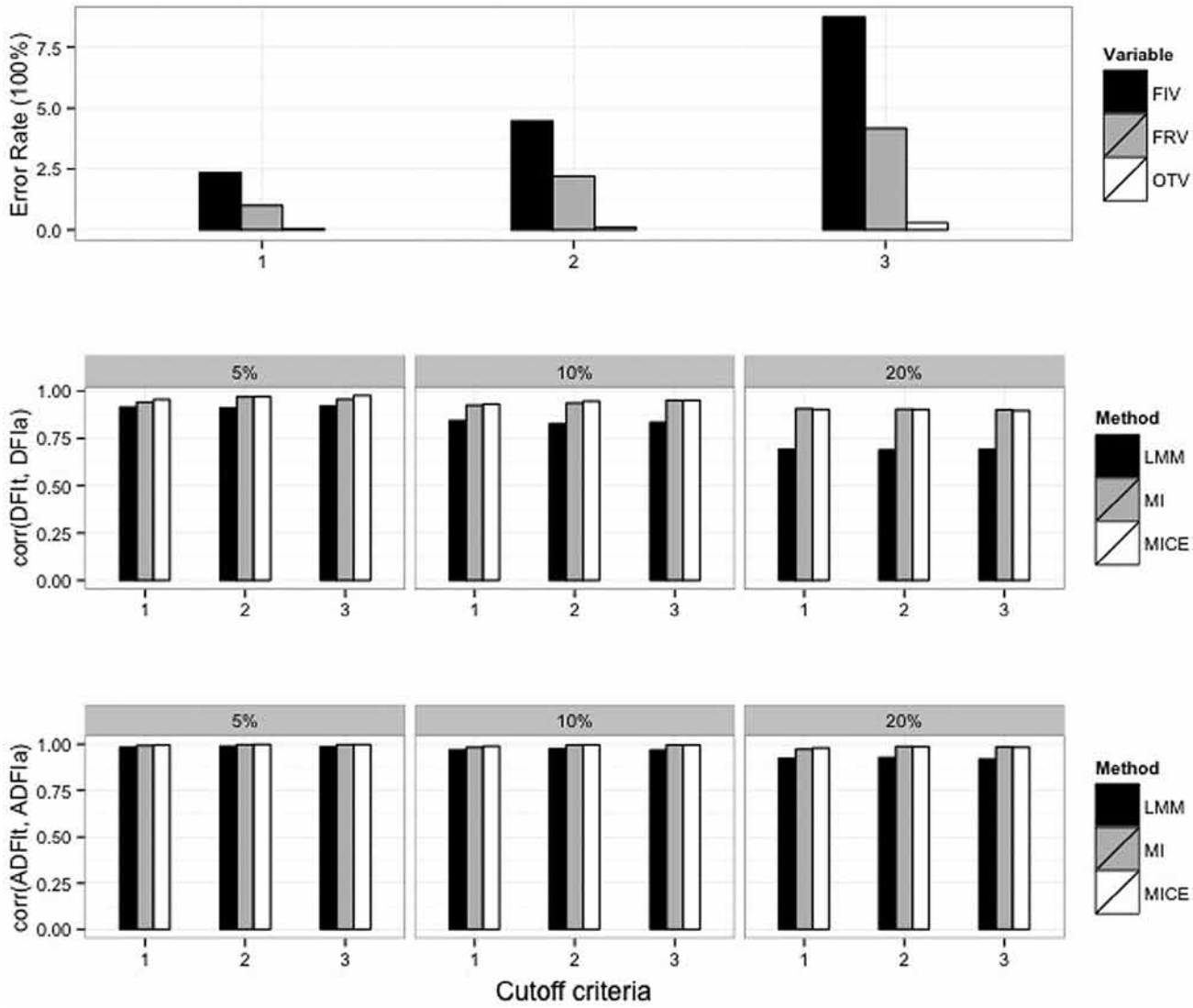

Figure 2. The impact of different cutoff criteria (for feed intake per visit [FIV] > 2,000 g for all visits; [-high], occupation time per visit [OTV] > 3,600 s for all visits [-high], and feed intake rate per visit [FRV] > 350 g/min for visits with FIV > 50 g [-high]) in the full data set on the performance of error-adjusting methods. Three different cutoff criteria (denoted as 1, 2, and 3 on the *x*-axis) were used based on the distribution of FIV, OTV, and FRV. The change of cutoff values impacted the 3 error types (FIV-high, OTV-high, and FRV-high) in the full data set (top figure), where cutoff criteria 2 is used in literature (Casey, 2003) and cutoff criteria 1 is half the cutoff and cutoff criteria 3 is double the cutoff in the right tail of their distributions. The middle and bottom figure show the change of performance of methods linear mixed model (LMM) and 2 alternative implementations of multiple imputation, denoted as MI (multiple imputation) and MICE (multiple imputation by chained equation). DFIt = true daily feed intake; DFIa = adjusted daily feed intake; ADFIt = adjusted DFIt; ADFIa = adjusted ADFI.

in which **y** is a vector of DFI_{ef}; **b** is a vector of estimated effects including the fixed coefficients of contemporary group, regression coefficients for off-test BW, coefficients for percentage of error of the 6 error types, regression coefficients for DFI for a certain error type ($\mathbf{DFI_e}$) of error type 4 and 5, and regression coefficients for daily occupation time for a certain error type ($\mathbf{OTD_e}$) of error type 1, 2, 7, and 8 (Table 1); **u** is a vector of animal effect assuming $\mathbf{u} \sim N(\mathbf{0}, \boldsymbol{\sigma}_u^2\mathbf{I})$; **e** is a vector of model residuals assuming $\mathbf{e} \sim N(\mathbf{0}, \sigma_e^2\mathbf{I})$; and **X** and **Z** are the corresponding design matrices. To reduce the bias caused by the extreme values in DFI_e and OTD_e (defined as $DFI_e < 0$ g, $DFI_e > 3{,}500$ g, $OTDe < 0$ s, and $OTD_e > 5{,}000$ s; Casey, 2003), those extreme values were removed before fitting them into the LMM. Linear coefficient estimates from this LMM were then used to adjust DFI_{ef} by adding an adjustment term, which was computed using the coefficient estimates for the corresponding error types.

Multiple Imputation

Multiple imputation was designed to tackle the general problem of replacing missing values in a data set. For this specific study, an error was defined as an extreme value of FIV, OTV, or RFV departing from its (unobserved) true value. To estimate the true value, the extreme value in the error visit can be viewed as a missing value problem and can be "filled" with the average of a set of plausible imputed values by multiple imputation.

Downloaded from https://academic.oup.com/jas/article/94/2/824/4701700 by Università degli Studi di Firenze user on 24 February 2022

Two multiple imputation methods were used. The first one used the R *mi* package (Su et al., 2011) to select conditional models for different variable types using regression models. The other method used the R package *mice* (van Buuren and Groothuis-Oudshoorn, 2011) using multivariate imputation by chained equations.

The fundamental idea of multiple imputation as implemented in both R packages is to use chained equation algorithms to deal with multivariate missing values. The principle of chained equation is based on drawing random samples from the conditional posterior predictive distribution of missing values under a particular Bayesian framework (Rubin, 2004). Using a simplified example, let us denote the response of a univariate sample $Y = (y_1, y_2, \ldots, y_n)$, in which the first values $Y_{obs} = (y_1, y_2, \ldots, y_a)$ are observed and the remaining values $Y_{mis} = (y_a + 1, y_a + 2, \ldots, y_n)$ are missing at random. Under an independent normal model $y_i \sim N(\mu, \varphi)$, i = 1, 2, …, n and $\theta = (\mu, \varphi)$ is an unknown parameter. The observed-data posterior distribution of θ with the uninformative standard prior $P(\theta) \propto \varphi^{-1}$ is

$$\mu|\varphi, Y_{obs} \sim N(\overline{Y_{obs}}, a^{-1}\varphi) \text{ and}$$

$$\varphi|Y_{obs} \sim (a\text{-}1)S^2_{obs}/\chi^2_{a-1},$$

in which $\overline{Y_{obs}} = a^{-1}\sum_{i=1}^{a} y_i$, $S^2_{obs} = (a-1)\sum_{i=1}^{a}\left(y_i - \overline{y_{obs}}\right)^2$, and χ^2_{a-1} denote a χ^2 variate with $a - 1$ df. To create an imputation $Y^{(1)}_{obs} = [y^{(1)}_{a+1}, \ldots, y^{(1)}_{n}]$, one would draw $y_i^{(1)} \sim N[\mu^{(1)}, \varphi^{(1)}]$ independently with a random mean $\mu^{(1)} \sim N[\overline{y_{obs}}, a^{-1}\varphi^{(1)}]$ and a random variance $\varphi^{(1)} \sim (a-1)s^2_{obs}/\chi^2_{a-1}$.

Repeating the procedure for l = 2, …, *m* results in m proper imputations for Y_{mis}.

More generally, $Y = (Y_{obs}, Y_{mis})$ following some parametric model $P(Y|\theta)$, in which θ has a prior distribution and Y_{mis} is missing. Because

$$P(Y_{mis}|Y_{obs}) = \int P(Y_{mis}|Y_{obs}, \theta)P(\theta|Y_{obs})d\theta.$$

an imputation for Y_{mis} can be created by first simulating a random draw of unknown parameters from their observed-data posterior $\theta^* \sim P(\theta, Y_{obs})$ followed by a random draw of the missing values from their conditional predictive distribution $Y^*_{mis} \sim P\left(Y_{mis}|Y_{obs}, \theta^*\right)$.

In both *mi* and *mice* packages (Su et al., 2011; van Buuren and Groothuis-Oudshoorn, 2011), it is assumed that the complete data Y is a partially observed random sample from the p-variate multivariate distribution $P(Y|\theta)$, assuming that the multivariate distribution of Y is completely specified by θ (the unknown parameters). To obtain a posterior distribution of θ, the chained equation algorithm proposes to iteratively sample from conditional distributions of the form $P(Y_1|Y_{-1}, \theta_1)$, …, $P(Y_p|Y_{-p}, \theta_p)$. Therefore, at the ith iteration, the chained equation is a Gibbs sampler, which draws

$$\grave{e}_1^{*(1)} \sim P\left(\grave{e}_1 \mid Y_1^{obs}, Y_2^{(i-1)}, \ldots, Y_p^{(i-1)}\right)$$

$$Y_1^{*(i)} \grave{e} P\left(Y_1 \mid Y_1^{obs}, Y_2^{(i-1)}, \ldots, Y_p^{(i-1)}, {}^{*(i)}_{1}\right)$$

$$\vdots$$

$$\grave{e}_p^{*(i)} \sim P\left(\grave{e}_p \mid Y_p^{obs}, Y_1^{(i)}, \ldots, Y_{p-1}^{(i)}\right)$$

$$Y_p^{*(i)} \grave{e} P\left(Y_p \mid Y_p^{obs}, Y_1^{(i)}, \ldots, Y_p^{(i)}, {}^{*(i)}_{p}\right)$$

in which $Y_j^{*(i)} = \left[Y_j^{obs}, Y_j^{*(i)}\right]$ is the jth imputed variable at iteration i. Because, in this way, the Gibbs sampler can be easily implemented as a concatenation of univariate procedures to fill out the missing data, this algorithm is called chained equation algorithm (van Buuren and Groothuis-Oudshoorn, 2011). As demonstrated in many studies (van Buuren et al., 1999**;** Rubin, 2003; Heymans et al., 2007), a low number iteration (say, 10 to 20) is often sufficient to carry out the imputation. By replacing the missing values with a set of imputed plausible values, multiple imputation generates

Table 2. Error rates in simulated replicated data sets

Replicate data set (Rep)	Error 1[1]	Error 2	Error 4	Error 5	Error 7	Error 8
Error rate 5%						
Rep 1	1.69	2.28	0.30	0.05	1.14	0.48
Rep 2	1.67	2.40	0.27	0.05	1.10	0.49
Rep 3	1.63	2.41	0.26	0.06	1.23	0.47
Rep 4	1.66	2.45	0.26	0.08	1.21	0.40
Rep 5	1.75	2.40	0.25	0.05	1.16	0.36
Mean	1.68	2.39	0.27	0.06	1.17	0.44
SD	0.04	0.06	0.02	0.01	0.05	0.06
Error rate 10%						
Rep 1	3.43	4.83	0.54	0.10	2.34	0.82
Rep 2	3.35	4.75	0.57	0.11	2.35	0.91
Rep 3	3.40	4.85	0.49	0.12	2.45	0.86
Rep 4	3.32	4.86	0.55	0.11	2.35	0.78
Rep 5	3.26	4.78	0.63	0.10	2.45	0.86
Mean	3.35	4.82	0.56	0.11	2.38	0.84
SD	0.07	0.05	0.05	0.01	0.06	0.05
Error rate 20%						
Rep 1	6.50	9.66	1.17	0.22	4.72	1.74
Rep 2	6.64	9.57	1.20	0.21	4.72	1.72
Rep 3	6.73	9.49	1.10	0.26	4.71	1.70
Rep 4	6.73	9.61	1.13	0.22	4.70	1.67
Rep 5	6.67	9.53	1.17	0.22	4.64	1.73
Mean	6.65	9.57	1.16	0.22	4.70	1.71
SD	0.10	0.07	0.04	0.02	0.03	0.03

[1]Error is indexed in this table and the unit is 100%. Error 1 = feed intake per visit (FIV) < –20 g for all visits (-low); Error 2 = FIV > 2,000 g for all visits (-high); Error 4 = occupation time per visit (OTV) < 0 s for all visits [–low]; Error 5 = OTV > 3,600 s for all visits [-high]; Error 7 = feed intake rate per visit (FRV) > 350 g/min for visits with FIV > 50 g [–high]; Error 8 = FRV = 0 g/min for visits with OTV > 500 s [-0].

Downloaded from https://academic.oup.com/jas/article/94/2/824/4701700 by Università degli Studi di Firenze user on 24 February 2022

Table 3. Accuracies[1] of adjusted daily feed intake (DFI_a) with 3 different error adjustment methods[2]

Replication data set	LMM			MI			MICE		
	5% rate[3]	10% rate	20% rate	5% rate	10% rate	20% rate	5% rate	10% rate	20% rate
1	90.70	82.33	70.63	96.78	92.17	90.48	96.77	94.58	90.40
2	90.50	81.37	66.89	96.37	92.89	89.76	96.83	93.89	90.20
3	90.75	82.89	69.80	97.05	94.59	90.32	96.50	94.79	90.81
4	91.63	83.60	67.69	95.88	94.81	90.81	97.06	93.75	89.62
5	91.38	82.99	68.45	97.49	92.78	89.84	97.03	95.10	89.63
Mean	90.99	82.64	68.69	96.71	93.45	90.24	96.84	94.42	90.13
SD	0.49	0.84	1.52	0.62	1.18	0.44	0.23	0.58	0.52

[1]Accuracies of DFI_a with 3 methods were evaluated with Pearson correlation coefficients of DFI_a and true daily feed intake (unit = 100%).

[2]Error adjustment methods include linear mixed model (LMM) approach and multiple imputation with MI (multiple imputation) and MICE (multiple imputation by chained equation).

[3]To obtain the simulated replication data sets, error visits were introduced to the "error-free" complete data set with 5, 10, and 20% rates.

multiple imputed data sets to reflect the uncertainty of imputed values, and statistical analysis needs to be appropriately applied to combine results obtained from each of them. For simplicity, the mean of the set of plausible values was used, because it may be viewed as the expectation of the imputed values for the unobserved entry in the data set.

Although R packages *mi* and *mice* are both based on the same chained equation algorithm, they use different elementary imputation methods to impute numeric missing values: predictive mean matching in package *mi* (Su et al., 2011) and Bayesian linear regression in *mice* (van Buuren and Groothuis-Oudshoorn, 2011). Predictive mean matching is a semiparametric imputation method that is restricted to the observed values and can preserve nonlinear relations in the conditional model whereas Bayesian linear regression is faster and more efficient when the residual of the conditional model is normal. For each simulated data set, the error visits identified and then masked as unobserved values (missing values) were drawn using Markov chain Monte Carlo as specified above. Convergence of the chains for both MI and MICE were examined by trace plots of the chains and assessed with the use of the CODA package in R (van Buuren and Groothuis-Oudshoorn, 2011; Su et al., 2011).

Measuring Method Performance

To compare the efficiency and accuracy of error adjustment, each method was applied to the same 15 simulated data sets and the results were compared using Pearson correlation of adjusted DFI (**DFI_a**) or adjusted ADFI (**$ADFI_a$**) with true DFI or the true ADFI computed using the EFD. The choice of the statistic for comparison was consistent with that proposed by Casey (2003) when comparing different feed intake editing strategies. In addition to Pearson correlation, the Spearman rank correlation coefficient, mean bias error, and root mean square error were also reported for DFI_a and $ADFI_a$. Daily feed intake was computed as sums of FIV by test day for each pig; ADFI was defined as the mean DFI across testing period for each animal.

RESULTS

Error rates for the 8 predetermined error types in the FD are reported in Table 1 and ranged from 0.0 to 4.46%. Feed intake per visit > 2,000 g for all visits (FIV-high; 4.46%) and FIV < –20 g for all visits (-low; 3.07%) were the 2 most common error types. The overall error rate (defined as the ratio of number of visits with at least 1 error over the total number of visits) was 9.28% in FD.

The error rates for the 8 error types in the 15 simulated data sets are shown in Table 2. The variation of error rates among replicates within the first scenario (containing 5% error visits) was small (SD of error rates ranged from 0.01 to 0.10%) whereas variability of error rates under scenario 2 (containing 10% error visits) and 3 (containing 20% error visits) were approximately 2 or 3 times the error rates under scenario 1 (5% error rates), which is expected by design.

Accuracies for DFI_a and $ADFI_a$ are shown in Table 3 and 4 for all method/scenario combinations. For all scenarios, MI and MICE performed similarly and outperformed LMM with higher accuracy in DFI_a (Table 3). For the 5% error rate scenario, all 3 methods performed well with correlations between true and adjusted DFI of 91.0, 96.7, and 96.9% for LMM, MI, and MICE, respectively. For moderate and high error scenarios (10 and 20%), multiple imputation methods (MI and MICE) were considerably more effective than LMM in terms of accuracy. Average accuracies were 82.6% for LMM and 93.5 and 94.4% for MI and MICE, respectively, in the 10% error rate scenario. Average accuracies were 68.7% for LMM versus 90.2 and 90.1% for MI and MICE, respectively, in the 20% error rate scenario. For ADFIa, the trend remained similar although the differences

Downloaded from https://academic.oup.com/jas/article/94/2/824/4701700 by Università degli Studi di Firenze user on 24 February 2022

Table 4. Accuracies[1] of adjusted ADFI ($ADFI_a$) with 3 different error adjustment methods[2]

Replication data set	LMM			MI			MICE		
	5% rate[3]	10% rate	20% rate	5% rate	10% rate	20% rate	5% rate	10% rate	20% rate
1	98.89	97.91	91.53	99.87	99.60	98.78	99.88	99.72	98.70
2	98.91	98.19	93.70	99.86	99.55	99.02	99.88	99.58	98.60
3	99.18	97.42	92.49	99.88	99.64	98.81	99.85	99.62	98.95
4	98.37	97.47	92.43	99.33	99.63	98.87	99.88	99.69	98.67
5	99.27	97.24	94.06	99.88	99.52	98.53	99.88	99.67	98.71
Mean	98.93	97.65	92.84	99.76	99.59	98.80	99.88	99.66	98.72
SD	0.35	0.39	1.03	0.24	0.05	0.18	0.01	0.06	0.13

[1]Accuracies of $ADFI_a$ with 3 methods were evaluated with Pearson correlation coefficients of $ADFI_a$ and true average daily feed intake (unit = 100%).

[2]Error adjustment methods include linear mixed model (LMM) approach and multiple imputation with MI (multiple imputation) and MICE (multiple imputation by chained equation).

[3]To obtain the simulated replication data sets, error visits were introduced to the "error-free" complete data set with 5, 10, and 20% rates.

were less marked, with the accuracies of LMM (ranging from 92.8 to 98.9%) consistently lower than the accuracies of MI and MICE (ranging from 98.7 to 99.9%; Table 4). We additionally computed the Spearman rank correlation between DFIa and true DFI (Supplemental Table S1; see the online version of the article at http://journalofanimalscience.org), the regression coefficient of DFIa against true DFI (Supplemental Table S2; see the online version of the article at http://journalofanimalscience.org), the mean bias error of DFIa against true DFI (Supplemental Table S3; see the online version of the article at http://journalofanimalscience.org), and the root mean square error of DFIa (Supplemental Table S4; see the online version of the article at http://journalofanimalscience.org). It should be pointed out that in all cases, LMM underestimated DFIa more as compared with MI or MICE (Supplemental Tables S2 and S3; see the online version of the article at http://journalofanimalscience.org). Identical trends were also found for ADFIa against true ADFI (unpublished data). When the same simulated scenarios were repeated with different cutoff criteria, results did not change. In all cases, MI and MICE outperformed LMM regardless of the cutoff criteria chosen (Fig. 2).

DISCUSSION

There has been considerable interest in feed intake, feed efficiency, and feeding behavior in livestock and much of that interest has centered on the ability to obtain reliable genetic/genomic predictions for these traits. A sizable body of literature has been produced on the application of genomic information in the prediction of feed intake–related traits (Fan et al., 2010; Do et al., 2013; Sahana et al., 2013) and feeding behavior (Onteru et al., 2011). However, much less attention has been dedicated to improving the quality of the data that is used in these analyses. Feed intake data collected using electronic feeding contain inaccuracies, which significantly hinder the quality of any downstream use of these data (de Haer et al., 1992; Casey et al., 2005). Filtering and/or editing of raw feed intake data collected by electronic systems is important first step of any genetic improvement program and cannot be overlooked. The percentage of error visits was moderately high, with an average of 9.28% errors visits. Previously reported error rates in feed intake collected by electronic feeders varied among different data sets. In a similar population of Duroc, Jiao et al. (2014) found the overall error rates ranging between 14 and 35%. Eissen et al. (1998) reported error visits representing 6% of the total 385,329 feeding visits for 250 pigs. Similarly, Casey (2003) reported percentages of identified error visits of 4.33, 5.62, and 18.74% for 3 different data sets with 863,590 total visits for 893 pigs, 290,073 total visits for 591 pigs, and 162,638 visits for 237 pigs, respectively. It is unsurprising that the occurrence of errors in feed intake during electronic recording varies among different data sets because the electronic feeding systems are placed under varying environmental conditions. This further highlights the need to develop robust error adjustment methods with existing electronic feed intake collecting systems or new equipment with more accuracy and less scrubbing.

The LMM approach developed by Casey (2003) is routinely used to adjust error visits in feed intake records but presents some limitations. The data processing before the actual mixed model application is cumbersome. For instance, the daily records need to be computed for visits with and without errors. Moreover, subjective constraints (in DFI_e and OTD_e) must be set on the model to limit the bias arising from influential or extreme values in the predictors, which results in removing a proportion of DFI records before fitting the models. Although improved mixed model approaches have been investigated to deal with influential observation (see, for example, Strandén and Gianola, 1999), the effectiveness of those models in feed intake data adjustment has not been assessed. Lastly, the correction for DFI using LMM does

Downloaded from https://academic.oup.com/jas/article/94/2/824/4701700 by Università degli Studi di Firenze user on 24 February 2022

not take the unidentified visits (missing values in feed intake) into account because of upfront removal of those visits. Conversely, the multiple imputation approach outlined in the current paper is very general and can be easily implemented in a variety of settings with minimal data preprocessing or ad hoc adjustments. Furthermore, by treating error records as missing, the extreme values have no effect on the MI and MICE models as compared with the LMM approach, where these have to be removed to ensure unbiased estimates. The results of our study provide a practical illustration of the advantages of MI or MICE over LMM in addressing the problem of occurrence of errors or outliers in feed intake data collected by electronic feeding systems.

The use of multiple imputation also has benefits with respect to the final data set after error adjustment. The main advantage is that multiple imputation produces a final data set with individual feeding records instead of one with individual ADFI across the whole testing period (see, for example, Eissen et al. [1999] and Hebart et al. [2004] for applications in swine and beef cattle). The LMM loses all the information of feed intake and related measures of individuals on a visit basis. Feed intake is a trait that intensely reflects the day-to-day or hour-to-hour dynamics of an animal's metabolism. To investigate the mechanisms of variation in individual feed intake over a testing period (Lorenzo Bermejo et al., 2003; Cai et al., 2011) or daily eating patterns (Young et al., 2011; Rohrer et al., 2013), feed intake and related measures such as feeding time or feeding rate per visit for each individual on test are required. In this situation, multiple imputation is a more natural method of choice because it allows one to make use of all the information provided.

In the statistical analysis of missing values, 3 mechanisms of missingness are considered (Rubin, 1976): missing completely at random (the probability of data being missing does not depend on the unobserved data or observed data), missing at random (**MAR**; the probability of data being missing does not depend on the unobserved data, conditional on the observed data), and missing not at random (**MNAR**; the probability of data being missing does depend on the unobserved data, conditional on the observed data). Feeding visits with errors or missing observations in feed consumption collected by electronic feeding systems result from animal–feeder interactions or feeder malfunction (Casey, 2003; Eissen et al., 1998). The probability of an error occurring does not appear, in this case, to depend on the amount of feed intake or time occupied of the feeder by the animal in that visit, given that electronic feeding systems record each visit of a pig to each feeder independently. Although multiple imputation can be implemented under MNAR (Carpenter et al., 2007; Rubin, 1976), standard implementations of multiple imputation assume MAR. We made the same assumption throughout this study. Future studies could relax this assumption.

Multiple imputation is a mature technique that is continually refined and makes it suitable for routinely handling missing data (Horton and Lipsitz, 2001). With the common inclusion of this program in statistical software, multiple imputation has become increasing popular to address erroneous and missing values, especially in medical and social science research (King et al., 2001; Royston, 2004; Sterne et al., 2009) to avoid bias for population parameter estimation in regression settings and loss of information due to missing values. In addition, multiple imputation is a very general data editing approach and could find broad applicability in all situations where error-prone data are used. Multiple imputation might be a proper technique to deal with errors or missing records in field-recorded data sets, such as disease incidence data from dairy producer–recorded health events from on-farm computer systems (Parker Gaddis, 2012).

Like any statistical techniques, it should be used after careful examination. As pointed out by Rubin (1996) and White et al. (2011), multiple imputation is not free of limitations and pitfalls. It is, for example, difficult to impute data points when the data set contains too many variables with missing values. Furthermore, the methodology is sensitive to error occurrence patterns and is computationally more intensive. It is implied that sensitivity analysis may be needed when applying multiple imputation to a new data set, and parallel computing might serve as a tool to release the computation burden.

In conclusion, we suggest multiple imputation as an effective alternative to LMM to deal with errors contained in feed intake data collected by electronic feeding systems. Application of multiple imputation in field data editing is exciting and encouraging and may need further investigation before use.

LITERATURE CITED

Allison, P. D. 2002. Missing data: Quantitative applications in the social sciences. Br. J. Math. Stat. Psychol. 55:193–196.

Cai, W., H. Wu, and J. C. M. Dekkers. 2011. Longitudinal analysis of body weight and feed intake in selection lines for residual feed intake in pigs. Asian-Australas. J. Anim. Sci. 24(1):17–27. doi:10.5713/ajas.2011.10142.

Carpenter, J. R., M. G. Kenward, and I. R. White. 2007. Sensitivity analysis after multiple imputation under missing at random: A weighting approach. Stat. Methods Med. Res. 16(3):259–275. doi:10.1177/0962280206075303.

Casey, D. S. 2003. The use of electronic feeders in genetic improvement programs for swine. PhD Diss., Iowa State Univ., Ames, IA.

Casey, D. S., H. S. Stern, and J. C. M. Dekkers. 2005. Identification of errors and factors associated with errors in data from electronic swine feeders. J. Anim. Sci. 83:969–982.

Downloaded from https://academic.oup.com/jas/article/94/2/824/4701700 by Università degli Studi di Firenze user on 24 February 2022

Chen, C. Y., I. Misztal, S. Tsuruta, W. O. Herring, J. Holl, and M. Culbertson. 2010. Influence of heritable social status on daily gain and feeding pattern in pigs. J. Anim. Breed. Genet. 127(2):107–112. doi:10.1111/j.1439-0388.2009.00828.x.

de Haer, L. C. M., J. W. M. Merks, H. G. Kooper, G. A. J. Buiting, and J. A. van Hattum. 1992. A note on the IVOG®-station: A feeding station to record the individual food intake of group-housed growing pigs. Anim. Prod. 54(01):160–162. doi:10.1017/S0003356100020717.

Do, D. N., A. B. Strathe, T. Ostersen, J. Jensen, T. Mark, and H. N. Kadarmideen. 2013. Genome-wide association study reveals genetic architecture of eating behavior in pigs and its implications for humans obesity by comparative mapping. PLoS One 8(8):e71509. doi:10.1371/journal.pone.0071509.

Eissen, J. J., A. G. De Haan, and E. Kanis. 1999. Effect of missing data on the estimate of average daily feed intake of growing pigs. J. Anim. Sci. 77(6):1372–1378.

Eissen, J. J., E. Kanis, and J. W. M. Merks. 1998. Algorithms for identifying errors in individual feed intake data of growing pigs in group-housing. Appl. Eng. Agric. 14:667–673. doi:10.13031/2013.19421.

Fan, B., S. Lkhagvadorj, W. Cai, J. Young, R. M. Smith, J. C. M. Dekkers, E. Huff-Lonergan, S. M. Lonergan, and M. F. Rothschild. 2010. Identification of genetic markers associated with residual feed intake and meat quality traits in the pig. Meat Sci. 84(4):645–650. doi:10.1016/j.meatsci.2009.10.025.

Hebart, M. L., W. S. Pitchford, P. F. Arthur, J. A. Archer, R. M. Herd, and C. D. K. Bottema. 2004. Effect of missing data on the estimate of average daily feed intake in beef cattle. Anim. Prod. Sci. 44(5):415–421. doi:10.1071/EA02109.

Heymans, M. W., S. van Buuren, D. L. Knol, W. V. Mechelen, and H. C. D. Vet. 2007. Variable selection under multiple imputation using the bootstrap in a prognostic study. BMC Med. Res. Methodol. 7:33. doi:10.1186/1471-2288-7-33.

Horton, N. J., and S. R. Lipsitz. 2001. Multiple imputation in practice: Comparison of software packages for regression models with missing variables. Am. Stat. 55(3):244–254. doi:10.1198/000313001317098266.

Jiao, S., C. Maltecca, K. A. Gray, and J. P. Cassady. 2014. Feed intake, average daily gain, feed efficiency, and real-time ultrasound traits in Duroc pigs: I. Genetic parameter estimation and accuracy of genomic prediction. J. Anim. Sci. 92(6):2377–2386. doi:10.2527/jas.2013-7338.

King, G., J. Honaker, A. Joseph, and K. Scheve. 2001. Analyzing incomplete political science data: An alternative algorithm for multiple imputation. Am. Political Sci. Rev. 95:49–69.

Lorenzo Bermejo, J., R. Roehe, V. Schulze, G. Rave, H. Looft, and E. Kalm. 2003. Random regression to model genetically the longitudinal data of daily feed intake in growing pigs. Livest. Prod. Sci. 82(2–3):189–199. doi:10.1016/S0301-6226(03)00032-0.

Onteru, S. K., B. Fan, M. T. Nikkilä, D. J. Garrick, K. J. Stalder, and M. F. Rothschild. 2011. Whole-genome association analyses for lifetime reproductive traits in the pig. J. Anim. Sci. 89(4):988–995. doi:10.2527/jas.2010-3236.

Osborne, J. W., and A. Overbay. 2004. The power of outliers (and why researchers should always check for them). Pract. Assess. Res. Eval. 9:1–12.

Parker Gaddis, K. L., J. B. Cole, J. S. Clay, and C. Maltecca. 2012. Incidence validation and relationship analysis of producer-recorded health event data from on-farm computer systems in the United States. J. Dairy Sci. 95(9):5422–5435. doi:10.3168/jds.2012-5572.

Rohrer, G. A., T. Brown-Brandl, L. A. Rempel, J. F. Schneider, and J. Holl. 2013. Genetic analysis of behavior traits in swine production. Livest. Sci. 157(1):28–37. doi:10.1016/j.livsci.2013.07.002.

Royston, P. 2004. Multiple imputation of missing values. Stata J. 4:227–241.

Rubin, D. B. 1976. Inference and missing data. Biometrika 63(3):581–592. doi:10.1093/biomet/63.3.581

Rubin, D. B. 1996. Multiple imputation after 18+ years. J. Am. Stat. Assoc. 91(434):473–489. doi:10.1080/01621459.1996.10476908.

Rubin, D. B. 2003. Nested multiple imputation of NMES via partially incompatible MCMC. Stat. Neerl. 57(1):3–18. doi:10.1111/1467-9574.00217.

Rubin, D. B. 2004. Multiple imputation for nonresponse in surveys. John Wiley & Sons, New York, NY.

Sahana, G., K. Veronika, H. Henrik, N. Bjarne, and O. F. Christensen. 2013. A genome-wide association scan in pig identifies novel regions associated with feed efficiency trait. J. Anim. Sci. 91(3):1041–1050. doi:10.2527/jas.2012-5643.

Sterne, J. A. C., W. R. Ian, J. B. Carlin, M. Spratt, P. Royston, M. G. Kenward, A. M. Wood, and J. R. Carpenter. 2009. Multiple imputation for missing data in epidemiological and clinical research: Potential and pitfalls. BMJ 338:b2393. doi:10.1136/bmj.b2393.

Strandén, I., and D. Gianola. 1999. Mixed effects linear models with *t*-distributions for quantitative genetic analysis: A Bayesian approach. Genet. Sel. Evol. 31(1):25–42. doi:10.1186/1297-9686-31-1-25.

Su, Y. S., M. Yajima, A. E. Gelman, and J. Hill. 2011. Multiple imputation with diagnostics (mi) in R: Opening windows into the black box. J. Stat. Softw. 45(2):1–31. doi:10.18637/jss.v045.i02.

van Buuren, S., H. C. Boshuizen, and D. L. Knook. 1999. Multiple imputation of missing blood pressure covariates in survival analysis. Stat. Med. 18(6):681–694. doi:10.1002/(SICI)1097-0258(19990330)18:6<681::AID-SIM71>3.0.CO;2-R

van Buuren, S., and K. Groothuis-Oudshoorn. 2011. MICE: Multivariate imputation by chained equations in R. J. Stat. Softw. 45:1–68.

White, I. R., P. Royston, and A. M. Wood. 2011. Multiple imputation using chained equations: Issues and guidance for practice. Stat. Med. 30(4):377–399. doi:10.1002/sim.4067.

Young, J. M., W. Cai, and J. C. M. Dekkers. 2011. Effect of selection for residual feed intake on feeding behavior and daily feed intake patterns in Yorkshire swine. J. Anim. Sci. 89(3):639–647. doi:10.2527/jas.2010-2892.

Downloaded from https://academic.oup.com/jas/article/94/2/824/4701700 by Università degli Studi di Firenze user on 24 February 2022