Using registry data to identify individual dairy cows with abnormal patterns in routinely recorded somatic cell counts

Henningsen, Maj Beldring; Reimert, Mossa Merhi; Denwood, Matt; Gussmann, Maya Katrin; Kirkeby, Carsten Thure; Nielsen, Søren Saxmose

Published in:
Journal of Theoretical Biology

DOI:
10.1016/j.jtbi.2023.111718

Publication date:
2024

Document version
Publisher's PDF, also known as Version of record

Document license:
CC BY

Citation for published version (APA):
Using registry data to identify individual dairy cows with abnormal patterns in routinely recorded somatic cell counts

Maj Beldring Henningsen*, Mossa Merhi Reimert, Matt Denwood, Maya Katrin Gussmann, Carsten Thure Kirkeby, Søren Saxmose Nielsen

Animal Welfare and Disease Control, Department of Veterinary and Animal Sciences, Faculty of Health and Medical Sciences, University of Copenhagen, Frederiksberg C, Denmark

ARTICLE INFO

Keywords:
Udder health
Cattle
Mastitis
Wilmink
Woods

ABSTRACT

Data from the Danish milk recording system routinely enter the Danish Cattle Database, including somatic cell counts (SCC) for individual animals. Elevated SCC can signal intramammary inflammation, suggesting subclinical mastitis. Detecting mastitis is pivotal to limit severity, prevent pathogen spread, and target treatment or culling. This study aimed to differentiate normal and abnormal SCC patterns using recorded registry data.

We used registry data from 2010 to 2020 for dairy cows in herds with 11 annual milk recordings. To create consistency across herds, we used data from 13,996 unique animals and eight different herds, selected based on the amount of data available, only selecting Holstein animals and conventional herds. We fitted log10-transformed SCC to days in milk (DIM) using the Wilmink and Wood’s curve functions, originally developed for milk yield over the lactation. We used Nonlinear Least Square and Nonlinear Mixed Effect models to fit the log10-transformed SCC observations to DIM at animal level.

Using mean squared residuals (MSR), we found a consistently better fit using a Wood’s style function. Detection of MSR outliers in the model fitting process was used to identify animals with log10(SCC) curves deviating from the expected “normal” curve for that same animal. With this study, we propose a method to identify single animals with SCC patterns that indicate abnormalities, such as mastitis, based on registry data. This method could potentially lead to a registry data-based detection of mastitis cases in larger dairy herds.

1. Introduction

Somatic cell counts (SCC) in dairy cattle are linked to intramammary infections, or mastitis, as an increased SCC level can indicate mastitis (Bradley & Green, 2005; Madouasse et al., 2010). In Denmark, mastitis is the main reason for the use of antibiotics in cattle (DANMAP, 2019; Kuipers et al., 2016). Mastitis is challenging to control, although the infections can severely affect animal welfare and milk production, the latter also affecting the economy of the farmers (Halasa et al., 2007; van den Borne et al., 2010).

SCC is measured routinely in Denmark in milk samples in the majority of dairy herds (about 89%), with most of these herds having 11 recordings per year (RYK Fonden, 2012). This allows us to monitor the SCC in individual animals, although without the ability to detect mastitis cases that may appear in-between observations. However, we can detect increased SCC level over time, which can be a sign of chronic mastitis (Nickerson, 2009). These chronic cases are important to detect as the infection may be able to spread within the herd from animal to animal, and these chronic cases may not always be treatable. However, to study the treatment efficacy, we must also be able to identify chronic cases.

To identify mastitis-caused fluctuations in the SCC, we must understand how to differentiate these from natural occurring fluctuation. Previous research has described fluctuations in the SCC caused by mastitis (De Haas et al., 2002, 2004; Deng et al., 2021), and it has also been shown that a known function for milk yield lactation curves, the Wilmink function (Wilmink, 1987), can be modified to fit the SCC to days in milk (DIM) (Grassbøll et al., 2016; Henningsen et al., 2023). However, a Wilmink style function showed to fit poorly (Henningsen et al., 2023). Therefore, given the well-established status of the Wilmink lactation curve in dairy science, it was of interest to compare its capabilities to fit SCC to DIM to another widely recognized curve function – the Wood’s curve function (Wood, 1967).

When we can identify mastitis-induced SCC fluctuations, we can
develop a warning system rooted in SCC for mastitis decision support. Furthermore, within the dairy industry’s shift towards data-driven decision making, as presented in the Dairy Brain study (Cabrera et al., 2020), our study endeavours to tap into routinely collected data. This data-driven approach enables farmers to take well-informed actions, including targeted treatment, additional diagnostic testing, and adherence to national regulations. These efforts are geared towards the overarching objectives of improving animal welfare and farmer profitability, while reducing antibiotic usage in dairy cattle. An additional advantage lies in the cost-effective nature of this approach, utilizing existing SCC registry data while giving farmers direct access to real-time SCC analysis from their herds.

Our aim was to propose and evaluate methods that can be used to identify infected animals, thus preventing potential circulation of mastitis pathogens in dairy cattle herds. The objective of this study was to use SCC registry data to identify individual animals with abnormal SCC levels from model fitting of larger data sets. We did this by fitting SCC to DIM and identifying the outliers visually in a mean square residuals (MSR) plot, the latter previously shown with pH data from dairy cattle linked to subclinical diseases (Denwood et al., 2018).

2. Materials and methods

2.1. Danish Cattle Database

We conducted an observational retrospective registry study, using registry data from the Danish Cattle Database (SEGES P/S, Aarhus N, Denmark). From this, we used five different data tables: The milk recording scheme, herd information, breed information, calving information, and animal information. Furthermore, we used treatment data for subsequent evaluation of the predicted SCC observations. We retrieved data covering a 10-year period from 2010 to 2020 from all Danish dairy herds with 11 milk quality controls per year, which is the standard in the milk recording scheme in Denmark, and is followed by approximately 90 % of herds. From the milk quality control data routinely collected through the Danish milk recording scheme, SCC in 10^3 cells per ml were obtained. SCC observations are limited to 1,000 cells per ml ≤ SCC < 9,999,000 cells per ml, since these two values constitute the lower and upper SCC limit. We used the date of calving to calculate DIM and parity to assign animals to parity groups 1, 2, 3, and >3. We further reduced data to only include Holstein cows in conventional herds, which constitute >77 % of Danish dairy cattle (Henningsen et al., 2023).

For practical reasons (to ease modelling feasibility), eight herds from 2,970 eligible herds were selected for further analysis. We selected these eight herds by identifying herds with the highest number of SCC observations over the 10-year data collection period. This allowed for a larger sample size from the eight herds, to enhance the stability of the modelling process. With more SCC observations, the model estimation becomes more robust, reducing the risk of convergence issues or numerical instabilities during the parameter estimation process. Then, only animals with a minimum lactation phase of 250 days and nine observations per parity were included. Nine SCC observations per animal per parity characterises almost 40 % of the animals in our data, and thus represents

![Fig. 1. Presentation of the process of data selection, modelling, assessment of the models and identifying animals with deviating somatic cell count (SCC) patterns. NLS = Nonlinear Least Square, NLME = Nonlinear Mixed Effect. Animals with high MSR are potentially those with mastitis.](null)
the broader population of herds adequately. We included SCC observations up to a lactation length of 305 days, this being the same DIM used to define the original Wilmink function for milk (Wilmink, 1987).

We present an overview of the data selection, modelling, and analysis in Fig. 1.

2.2. Summary statistics

We calculated SCC summary statistics for each herd and for all the individual animals. We calculated the mean, median, and the first and third quantile. We used R for all data handling (R Core team, 2020).

2.3. Statistical methods for SCC data

To fit SCC to DIM we used two different modified lactation functions, Wilmink and Wood’s. The original Wood’s lactation function is fitted with log10(SCC) as outcome and modelling DIM with three parameters (Wood, 1967). We derived a three parameter Wood’s style function to fit the log10-transformed SCC to DIM:

\[
\log_{10}(SCC) = \log(a) + b \cdot \log(DIM) - (c) \cdot DIM
\]

(1)

where the subscript \( j \) represent the random slope of the \( j \)th animal, and \( a \), \( b \), and \( c \) are the Wood’s style parameters describing the log10(SCC) curve. Note that the random slope for animal allows the parameters to vary between animals, although these will also be influenced by the average parameter values for all animals in that herd due to shrinkage.

From the original Wilmink lactation function (Wilmink, 1987), we derived the following Wilmink’s style function:

\[
\log_{10}(SCC) = d \cdot \exp\left(\frac{f \cdot DIM}{c}\right)
\]

(2)

where the subscript \( j \) represent the random slope of the \( j \)th animal, and \( d \), \( e \), and \( f \) are the Wilmink style parameters describing the log10(SCC) curve.

We used a nonlinear least square (NLS) model and a nonlinear mixed effect (NLME) model to fit SCC to DIM on herd and animal level, refitting the model to each herd separately, using both the Wood’s style function and the Wilmink style function (Henningsen et al., 2023). Random slopes of animals were fitted to each of the three parameters in both functions, in order to allow for variation in SCC curve between animals (Henningsen et al., 2023).

An NLS model finds the best fit of a nonlinear model to a set of data points by minimizing the sum of the squares of the differences between the observed data points and the model (Bates & Chambers, 1991). The NLME model combines nonlinear regression with random effects, thus allowing fitting a model with both fixed and random effects (Pinheiro et al., 2021). The use of random effects allows for the incorporation of animal variation, while fixed effects link different animals together in the analysis ( Laird & Ware, 1982). Both NLS and NLME can be used to fit SCC to DIM; however, only the NLME model will account for correlations between SCC observations.

For NLS, we used the nlstools function in R (Padfield & Matheson, 2020). Our purpose with fitting to NLS was to create functional initial parameter values for NLME. NLS was fitted on population level, while NLME was fitted on herd level and with random effects on animal level, using the NLME function in R with maximum log-likelihood optimization (Pinheiro et al., 2021).

2.4. Model assessment

We used mean squared residuals (MSR) to evaluate the overall goodness-of-fit of our models. The residuals describe the differences between the observed SCC values and the model predicted SCC values. Squared residuals allowed us to favour small errors over large errors, since the squared value of a larger error is much greater than the squared value of a smaller error. We compared our Wilmink style function to Wood’s style function, identifying the model with the lower MSR, as it indicates that the model has a better fit to the data (given that both functions have the same number of parameters). We then evaluated how well that model fitted for each herd and parity, and for each DIM, grouping DIM in groups of five days, to simplify the figure.

We also conducted a case study involving four animals, one selected from each parity group, based on the highest MSR values observed from the Wood’s style function. The MSR values indicate the degree to which the animals’ SCC patterns deviated from what was expected, assuming the Woods style function mimics SCC observations from animals without mastitis. By specifically choosing the animals with the most extreme MSR values, our aim was to gain insights into the factors contributing to these pronounced deviations. It is important to note that while this case study approach focuses on a selected subset of animals, rather than providing a representative sample of the entire population, analysing these extreme MSR cases offers valuable insights into the potential underlying mechanisms. As a supplementary, animals with a MSR above a 95% confidence interval (CI), within the CI, and below CI, where randomly chosen and compared.

3. Results

3.1. Descriptive analysis

We used a total of eight herds and 13,996 unique animals in the study, with SCC observations recorded from 2010 to 2020. The independent variable, SCC, is presented in Table 1 and Table 2, where Table 1 presents a summary of SCC across the eight herds, and Table 2 presents a summary of SCC across the four parity groups. For the herds, across parities, we see Herd 2 had an overall higher SCC, with the 75% quantile surpassing a SCC level of 200,000 cells/ml; an often-used SCC threshold, to indicate mastitis (Olde Riekink et al., 2007; Zeconzi et al., 2020). We observed an increase in SCC, across herds, when the parity group increases, where both the parity groups 3 and 3 surpassed a SCC level of 200,000 cells/ml at the 75% quantile.

3.2. Somatic cell count modelling

We retrieved MSR from the NLME model fits. When comparing the Wood’s style function with the Wilmink style function using MSR from the respective NLME model fits (Table 3 and Fig. 2), we overall found a better fit using Wood’s rather than Wilmink, as summaries of MSR were lower for Wood’s than Wilmink.

Fig. 3 shows the MSR for Wood’s function grouped by herd and parity group, visualizing in which parity group and herd the lowest and highest MSR values were found. We see a difference in how well animals from each herd fitted to the models, with animals from Herd 4 achieving the best fit, while animals from Herd 8, with the highest MSR value, fitted the worst (Table 4). We also see, for all eight herds, that a better fit was associated with a lower parity group (Table 5). To evaluate model fitting with the Wood’s style function for each DIM, we plotted the NLME residuals over DIM, shown in Fig. 4. From this we see the largest

<table>
<thead>
<tr>
<th>Herd</th>
<th>Total SCC observations</th>
<th>Geometric mean</th>
<th>Min SCC</th>
<th>Q1 SCC</th>
<th>Median SCC</th>
<th>Q3 SCC</th>
<th>Max SCC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>13,608</td>
<td>73</td>
<td>1</td>
<td>34</td>
<td>62</td>
<td>132</td>
<td>9,136</td>
</tr>
<tr>
<td>2</td>
<td>25,992</td>
<td>101</td>
<td>1</td>
<td>41</td>
<td>81</td>
<td>202</td>
<td>9,892</td>
</tr>
<tr>
<td>3</td>
<td>31,473</td>
<td>81</td>
<td>1</td>
<td>33</td>
<td>65</td>
<td>153</td>
<td>9,967</td>
</tr>
<tr>
<td>4</td>
<td>25,011</td>
<td>62</td>
<td>1</td>
<td>28</td>
<td>55</td>
<td>125</td>
<td>9,802</td>
</tr>
<tr>
<td>5</td>
<td>35,424</td>
<td>78</td>
<td>1</td>
<td>31</td>
<td>66</td>
<td>174</td>
<td>9,840</td>
</tr>
<tr>
<td>6</td>
<td>25,353</td>
<td>86</td>
<td>1</td>
<td>35</td>
<td>72</td>
<td>180</td>
<td>9,948</td>
</tr>
<tr>
<td>7</td>
<td>22,797</td>
<td>77</td>
<td>1</td>
<td>27</td>
<td>62</td>
<td>184</td>
<td>9,786</td>
</tr>
<tr>
<td>8</td>
<td>20,925</td>
<td>65</td>
<td>1</td>
<td>26</td>
<td>53</td>
<td>127</td>
<td>9,981</td>
</tr>
</tbody>
</table>
spread of residuals around 50 DIM.

Predicted log10(SCC) curves for four animals identified as outliers with a high MSR value are shown in Fig. 5. The curves from the four animals, one from each parity group, all deviated from what we would expect of a normal log10(SCC) curve. A 'normal' log10(SCC) curve typically follows a pattern of gradual decrease, reaching a minimum, and then gradually increasing. This pattern reflects the typical SCC dynamics during a lactation phase in the absence of mastitis. An illustrative example of this typical curve is provided in Fig. 5, where it is depicted as a dashed curve plotted alongside each of the four high MSR animals. It serves as a reference, selected amongst the five animals with lowest MSR, representing a predicted log10(SCC) curve for an animal with low MSR. The log10(SCC) curves from the animals with high MSR in parities 1 and 2 were opposite to what we expected, with first a rise in SCC followed by decreasing SCC values towards the end of the lactation phase. The high MSR animals in Parity 3 and >3 predict SCC to follow a roughly straight line, without any local extrema created by SCC fluctuations. A supplementary figure, Fig. S.1, show curves from pairwise randomly selected animals.

### 4. Discussion

With this study, we have suggested a method to identify single animals for follow-up investigations on mastitis using registry data, specifically milk quality recordings, from dairy cows (combined with essential animal and herd information such as breed, parity, and herd type). The methods suggested could potentially be used in bigger herds for SCC surveillance of dairy cows, to identify animals for further characterization with regards to mastitis control.

Previously, Græsbøll et al. (Græsbøll et al., 2016) suggested a four-parameter Wilmink style function. We anticipated that a four-parameter function would generally yield a lower MSR compared to a three-parameter function. The reasoning behind this expectation was that with more degrees of freedom, the four-parameter function could potentially accommodate more complex patterns in the data (Friedman et al., 2001). Nonetheless, we still aimed to compare simpler functions, as our objective was to propose methods for mastitis surveillance that could be applied not only in individual herds with limited data but also in larger datasets encompassing entire populations, such as registry-based datasets. Thus, we decided to derive three-parameter functions from both original lactation functions. By selecting the most appropriate three-parameter function, we aimed to strike a balance between model simplicity and effective mastitis surveillance across various data scenarios.

We chose to use an NLME model to fit our SCC to DIM. NLME is sensitive to the initial parameter values when fitting (Pinheiro et al., 1995). This means that it may either not converge, or we can risk convergence to a local minimum, rather than a global (Pinheiro & Bates, 2001).
We increased the likelihood of finding the global minimum by first using an NLS model to fit SCC to DIM, using the NLS parameters as initial parameter values for the NLME model fitting (Henningsen et al., 2023). While the NLS model is less sensitive to the initial parameter values, random and fixed effects are not possible to fit (Laird & Ware, 1982), which is desirable when analysing animals and/or herds, where variations often occur. We therefore chose an NLME model, rather than NLS, for our main analysis.

From our results (Fig. 4), we observed how the goodness of fit varied with DIM, where the largest spread of residuals was around 50 DIM. This may be a result of having relatively few observations to describe the quick initial drop on the SCC curve, followed by the SCC minimum, and the first part of the SCC increase after the minimum. These events all take place within 1–100 DIM, where we expect a maximum of three SCC observations per animal, whereas on the increasing slope after the minimum, more information is available, often with a minimum of six SCC observations.

From the Wood’s style function fit, we selected outliers with high MSR, from each parity group. To understand what could have caused these animals to fit poorly, we retrieved registry-based treatment in-formation that we could compare with their predicted log10(SCC) curves. The predicted log10(SCC) curves for animals with high MSR all differ from an animal with low MSR such shown in Fig. 5. For the high MSR animal in Parity 1 we see a low first SCC observation followed by high SCC observations, where we otherwise would expect to meet the SCC minimum. This is followed by, at 150 DIM, a drastic SCC drop, where we would expect the SCC values to increase. This SCC behaviour may be explained by several treatments for endometritis in the beginning of the lactation phase before DIM 83. However, we do not find any treatments associated with the drop around 150 DIM. For the animal in Parity 2, we also did not see a minimum of SCC where expected. Instead, we saw high SCC observations throughout the lactation phase, except the last observation, where a SCC minimum (1000 cells/ml) was reached. For this animal we again identified endometritis treatments in the beginning of the lactation, but also a treatment for limb disorder in the middle of the lactation. No treatments were registered in the end of the lactation to explain the minimum of 1000 cells/ml. Given the extremely low value, it seems unexplainable, potentially indicating a recording error. Such an anomaly might result from sample mix-ups on the farm or incorrect SCC documentation, an observation corroborated by Fig. S.2 in our raw data. This supplementary figure provides confirmation, revealing 268,582 instances of SCC recordings with a value of 0, an unlikely low value. The animal in Parity 3 in Fig. 5 had been treated several times for mastitis from DIM 102 to 121, which could explain both the high SCC values around 100 DIM and the low SCC values in the beginning of lactation.
last phase of the lactation. And for the animal in a Parity > 3, treatments for acute mastitis around DIM 180 can both explain the high SCC peak followed by a drop towards the end of the lactation phase. What all these four animals, with a high MSR, have in common is, that they have all been diagnosed or treated in the lactation phase, which may have affected the SCC values. We did not find any treatments for the low MSR animal in Fig. 5 throughout the lactation.

The conditions we chose for our eligible data are a limiting factor. We chose eight predominantly larger herds, and only animals with exactly nine SCC observations in the lactation were retrieved for a minimum of 250 DIM. This enabled us to compare model fitting for a rather uniform sample population. However, when applying this model on the entire Holstein population of conventional Danish dairy cattle, we would expect limitations regarding the NLME model fitting with large data as described earlier. That is, the high sensitivity towards initial start parameters being further challenged with greater variation in the data. Also, we have a selection bias in this study. The eight herds with the maximum number of observations may not be representative of the broader population of herds. By focusing solely on herds with high observation counts, we may inadvertently introduce a bias towards certain herd characteristics or management practices. It is therefore worth noting that our method predominantly detects deviations from the ‘normal log10(SCC) curve’ when an adequate number of SCC measurements are included in the model, as exemplified by the requirement of nine SCC observations in this study. Thus, the efficacy in identifying mastitis cases may depend on the availability of a robust dataset with sufficient data points to establish a dependable baseline. This limits the generalizability of our findings to the entire population of herds, and therefore it is important to test this concept for more herds before roll-out to herds with less data. Focusing on herds with the maximum observations assumes that sufficient data are available for those herds. However, some herds may have missing or incomplete data, which can introduce bias or reduce the overall reliability of our analysis.

In addition to the limitations mentioned above, it is also important to acknowledge that we have not yet tested our methods with a real-time approach, meaning we have not evaluated if animals can be detected during the lactation phase based on their MSR values. This aspect remains unexplored and could be a valuable avenue for future research.

5. Conclusion

We have shown that a Wood’s style function can be used to fit registry SCC data to DIM, using an NLME model, and that Wood’s provides a better fit than a Wilmink style function. We have also shown that by catching MSR outliers from a Wood’s fit with an NLME model, we can identify animals with abnormal SCC levels, or abnormal SCC fluctuations, in the lactation phase compared with natural occurring SCC fluctuations throughout the lactation phase. This is supported by our finding that lower parity animals have lower MSR, which we suggest corresponds to a lower probability of having mastitis. This suggests that animals with a higher MSR may be further scrutinised for further testing, treatment, or culling, while considering a high MSR could also indicate treatment, but in such case the animal will already be under mastitis observation.

Overall, our findings emphasize the value of utilizing the Wood’s style function in conjunction with MSR analysis to identify and take appropriate action regarding animals displaying abnormal SCC dynamics during the lactation phase.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors would like to thank SEGES for have given access to the Danish Cattle Database. The Danish Veterinary and Food Administration (FVST, Glostrup, Denmark) funded this project but did not take part of the planning and conduct of the project.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jtbi.2023.111718.

References


Fig. 5. Somatic cell count (SCC) curves for four dairy cows identified as upper outliers using the mean squared residuals (MSR) calculated from fitting SCC to days in milk using a modified Wood’s lactation function. All curves identify animals with an abnormal SCC throughout the lactation phase. The dashed curves are fitted for a single animal in Parity 1 with low MSR. This reflect what would be expected of an animal without mastitis, with SCC never exceed a log10 SCC value of 5.3, corresponding to 200,000 cells/ml, a threshold often used to deem animals positive for mastitis.