Retrospective analysis of calving interval data from New Zealand dairy herds

Analysis log

[1 Plan 1](#_Toc155357878)

[2 Data collation 2](#_Toc155357879)

[2.1 Descriptive analysis 4](#_Toc155357880)

[2.2 Association between serum and BM Se 12](#_Toc155357881)

[3 Blood and bulk milk Se data – take 2 13](#_Toc155357882)

[3.1 Data collation 13](#_Toc155357883)

[3.2 Descriptive analysis 15](#_Toc155357884)

[3.3 Association between serum and BM Se 21](#_Toc155357885)

[4 Questionnaire 22](#_Toc155357886)

[5 Associations between the questionnaire and selenium concentrations 25](#_Toc155357887)

[6 References 26](#_Toc155357888)

# Plan

* ~~Set up R project, synch with GitHub, and set environment with renv,~~
* Import raw data and check and clean it,
* Descriptive analysis,
* Adjusted analysis, if necessary,
* Write report.

The objectives of this analysis are to:

1. Describe calving intervals for many VetEnt farms (in accordance with agreed inclusion and exclusion criteria),
2. Explore if there are differences in calving intervals between cows of different ages, production levels, and other factors to be defined.

# Data collation

Files created/downloaded:

|  |  |
| --- | --- |
| **Name** | **Description** |
| REPORT2019.xlsx | Raw data created by Nic Wilson ex Infovet – 19/20 season |
| REPORT2020.xlsx | Raw data created by Nic Wilson ex Infovet – 20/21 season |
| REPORT2021.xlsx | Raw data created by Nic Wilson ex Infovet – 21/22 season |
| REPORT2022.xlsx | Raw data created by Nic Wilson ex Infovet – 22/23 season |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

* There were 4 spreadsheets of raw data:
  + 19/20 582,662 observations,
  + 20/21 576,478 observations,
  + 21/22 540,734 observations,
  + 22/23 520,152 observations,
  + A total of 2,220,026 observations from 539,801 (277,748 using tag number) unique animals on 368 farms.
  + 1,148,535 observations are active, 1,071,491 inactive.
  + 377,231 observations lack a calving date, only 26,736 of which are males.
  + 1,838,543, 4,143, 100, 6, and 3 observations had 1-5 unique calving dates respectively. 24,351 observations with >1 calving date (not necessarily same or different). 4,252 observations with >1 unique calving date.
* Initial data cleaning
  + Remove males (n=26,736 observations).
  + Remove animals with >1 calving date in a season (n=24,351 observations).
  + Transcription error handling:

|  |  |  |  |
| --- | --- | --- | --- |
| **Farm** | **DCF1** | **Gribbles** | **Action** |
| Adam Joyce | 76 | 26 | Sas has taken 26 and inserted age from 76 |
| Brett Dann | 778 | 777 | Sas has taken 777 and inserted age from 778 |
|  | 864 | 866 | Sas has taken 866 and inserted age from 864 |
| Dwayne Little | 117 | 117 | These cannot be discerned, ages are different, so keep Se but leave age NA |
|  | 177 | 117 duplicate | These cannot be discerned, ages are different, so keep Se but leave age NA |
| Mikita van der Heyden | 18 | 8 | Sas applied the DCF1 ID as there is no 8 in DCF1 and used 18’s age |
| Kakariki Landscapes | 122 | 221 | DCF1 does not have a 221, but does have a 122, so Sas used the information for 122 but used 221 as the ID. |
| Farm E | 277 | 677 | DCF1 does not have a 677, but does have a 277, so Sas used the information for 277 but used 677 as the ID. |
| Farm M | 25 | 22 | DCF1 does not have a 22, but does have a 25, so Sas used the information for 25 but used 22 as the ID. |
| Farm B | 144 | 141 | DCF1 does not have a 141, but does have a 144, so Sas used the information for 144 but used 141 as the ID. |

* + Age is missing from 4 cows
    - 117 x 2 as above.
    - No tag on Farm H.
    - Unknown age on farm U.
  + All unique rows.
  + Renamed supply number “sn”.
  + 60 farms (supply numbers), 7 regions.
  + 8, 10 or 11 bloods/farm, 3-11 farms/region.
    - 39655 had 8 bloods,
    - 45007, 45627, and 45632 had 11 bloods.
    - These are all legit and what the vets collected according to Sas.
  + Sampled 20 Sep to 04 Oct.
  + Age = 2-13, median = 4
  + Blood Se ND, mean = 1343, SD=554, range = 55-2596, some regional variation, large between-farm variation.
  + Se thresholds:
    - 7/601 (1%) of observations had blood Se < 250 nmol/L, all on 1 farm (45636 - Manawatu), which has the lowest mean blood Se.
    - 41/601 (7%) were below 500.
    - 177/601 (30%) were below 1000.
* BM dataset:
  + 62 observations of 3 variables, all unique (there were another 6 variables that I dropped as they were not germane according to Sas).
  + 60 unique supply numbers - 2 farms have 2 observations on same date but different Se values:
    - 36843 (Makaiwai, Canterbury, 26/09/2023).
    - 74570 (Jaden Gray, Otorohanga/TA, 02/10/2023).
    - Duplicates were highlighted in the raw data (but not in Sas’ sheet), so the duplicates were filtered out.
    - Only 35 unique values of Se, which are very low values without great precision.
  + BM Se is NOT ND (right tail), median=0.013, range = 0.0036-0.0380
  + BM and serum dates generally disagree – up to 6 days out.
* Questionnaire data
  + 60 observations, 85 questions.
  + Breed (1.5) categorised into F, XB and J, where F = anything with Friesian and no other breed, or Friesian-FX (n=1).
  + Milk production (1.6): took midpoint of range, multiplied cow-level values by N cows at peak, values expressed per day were set as NA.
  + NICR (1.8): midpoint of ranges taken.
  + SCC: midpoint of ranges taken.
  + Supplementation (2.1): 4/5 of the “Sometimes – depends” answers were in fact “Yes” – change to Yes. The one “Sometimes” farmer stated “Maybe in mineral mix”.

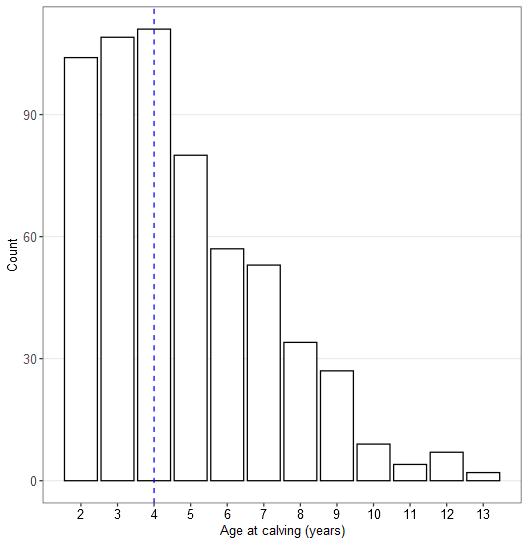
## Descriptive analysis

* 601 observations of 10 variables:
  1. Supply number
  2. Farm name
  3. Region
  4. Cow ID
  5. Blood sample date
  6. Cow age
  7. Blood Se
  8. Bulk milk test date
  9. Bulk milk Se
  10. Study ID
* Age is missing from 4 cows: 2 cows could not be differentiated from each other; 2 cows were missing age in raw data.
* All unique.
* 60 farms (supply numbers), 7 regions, 3-11 farms/region (table).

**A white paper with black text

Description automatically generated**

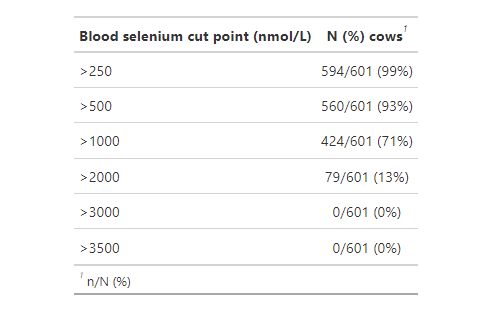
* 1 farm had 8 bloods (Canterbury), 56 had 10, and 3 had 11 bloods (Manawatu).
* Sampled 20 Sep to 04 Oct.
* Age = 2-13, median = 4, not as many heifers as expected



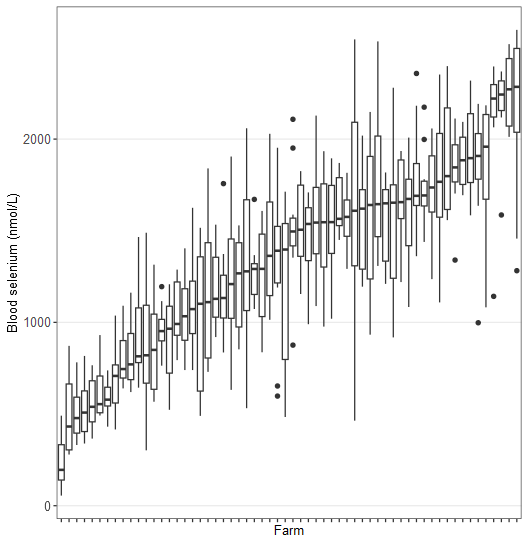
* + Blood Se ND, mean = 1343, SD=554, range = 55-2596, some regional variation, large between-farm variation.

A graph of a number of columns

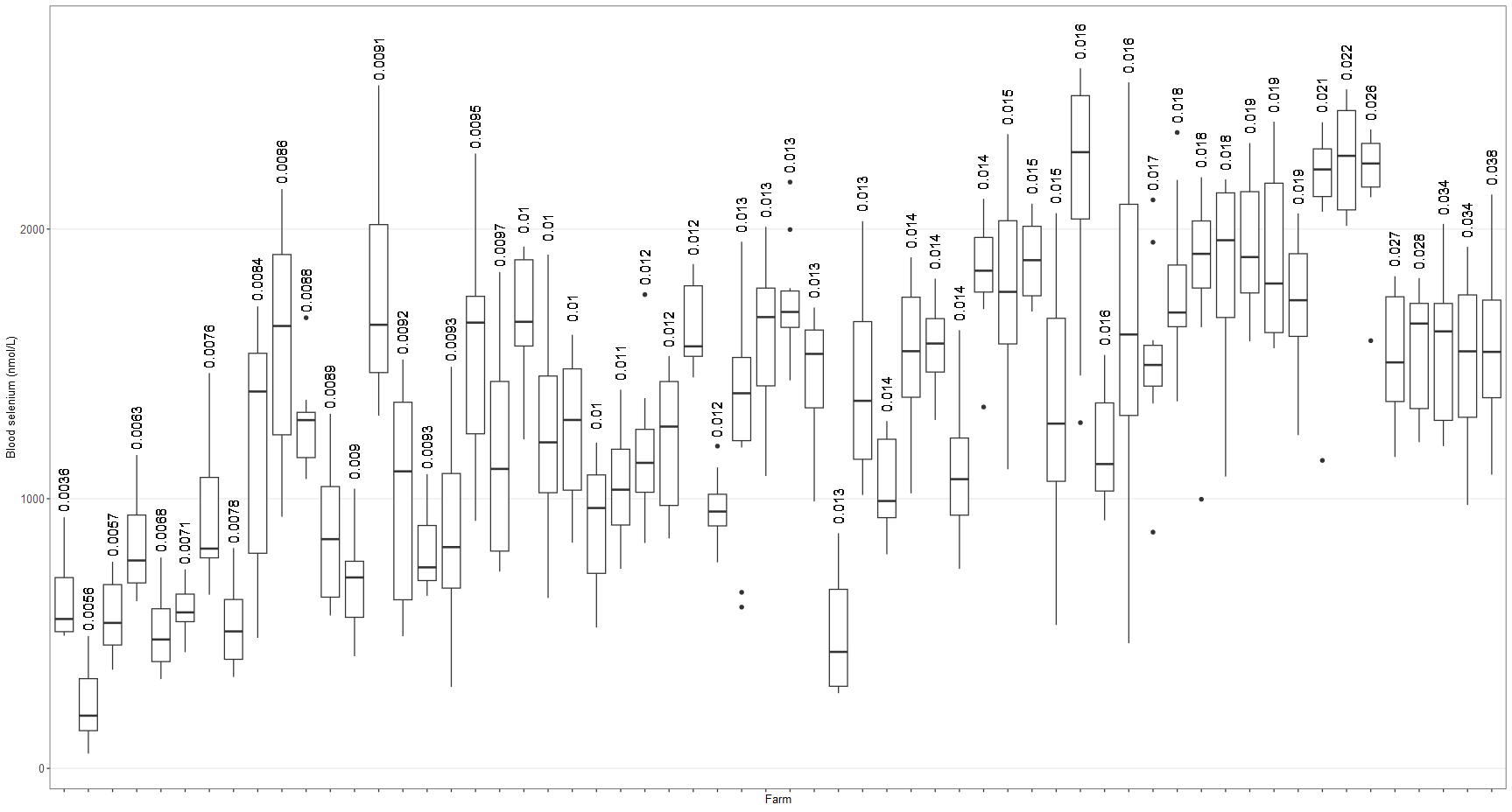
Description automatically generated



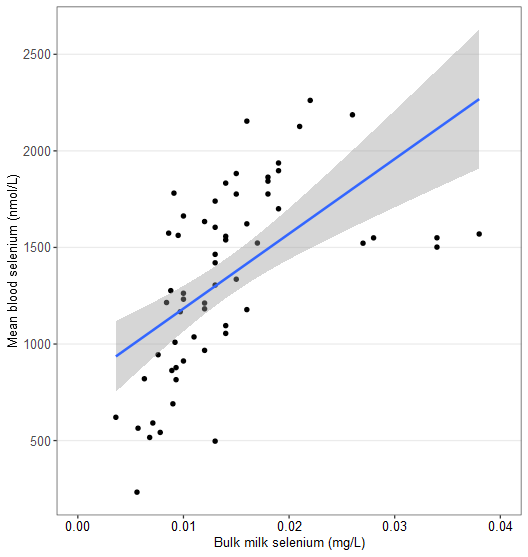
* Quite a bit of variation within and between farms in blood Se



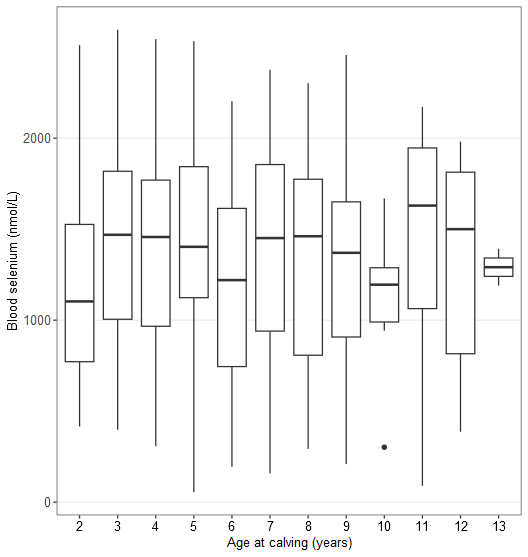
* When sorted on bulk milk Se, there is clearly an overall positive association, but it shows the large within-farm variance in blood Se.



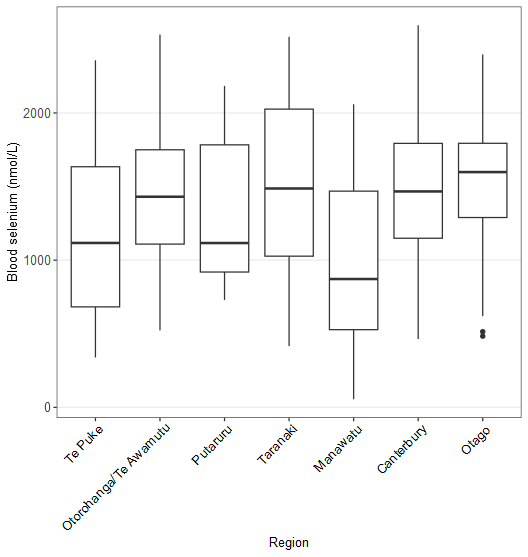
* When BM Se is plotted against mean blood SE, the positive relationship is clearer, but there are some high BM Se farms with lower than expected blood Se. 4 are from Canterbury, 1 from Taranaki. Maybe large herds?



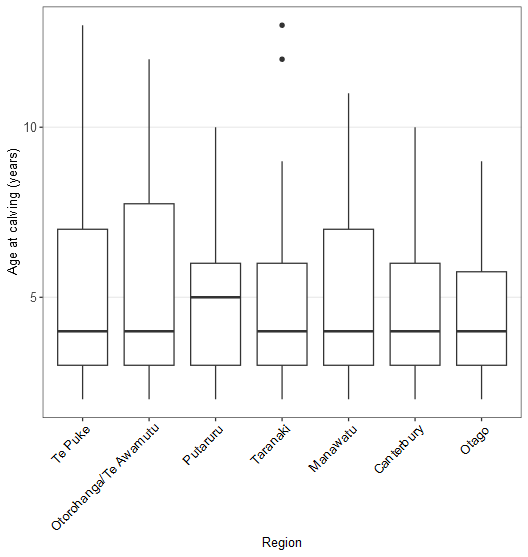
* No clear association with age:



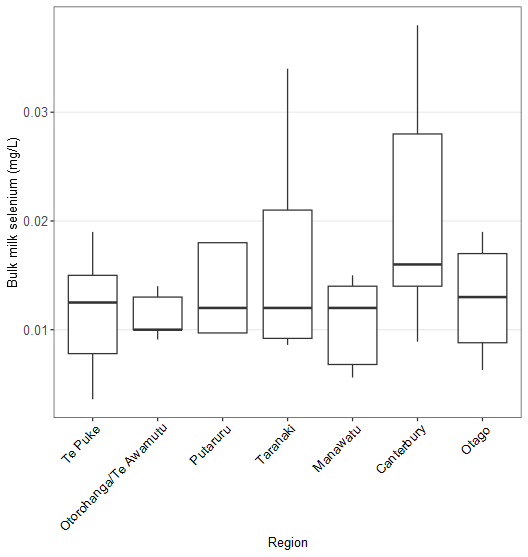
* No clear association with region, though Manawatu was lower:



* Age doesn’t seem to be associated with region:



* BM dataset:
  + 62 observations of 3 variables, all unique (there were another 6 variables that I dropped as they were not germane according to Sas).
  + 60 unique supply numbers - 2 farms have 2 observations on same date but different Se values:
  + BM Se is NOT ND (right tail), median=0.013, range = 0.0036-0.0380
  + BM and serum dates generally disagree – up to 6 days out.
  + No clear association between BM Se and region: though Canterbury seems high and Manawatu low:



## Association between serum and BM Se

* I scaled BM Se by multiplying it by 1000 (converting mg to g) to show the effect of a 1mg increase in BM Se instead of a 1g increase, which makes no sense.
* Simple linear models, no other variables because I am taking mean blood Se, so age is irrelevant.
* BM Se led to a non-linear model. Adding a square term improved fit (LR test and AIC) and resolved nonlinearity. Adding a cubed term made no improvement.
* Diagnostics
  + Residuals are ND, homoscedastic, and linear.
  + Farms 21659 and especially 36765 have large leverage and CDs due to lowest and highest BM Se respectively, but not as extreme blood Se.
  + Farm 45632 had the highest residual (lower blood Se than expected).

# Blood and bulk milk Se data – take 2

Here, I need to use the same dataset as take 1, but add the BM and blood samples collected at the second visit from the 5 farms that were suspect at the first visit (36765, 36770, 36843, 37105, 42594).

## Data collation

Sas has created new files that contain all data from both visits. I will just use the new files and:

* Discard the first blood sample for the 5 farms from the blood dataset,
* Discard the second milk sample from all farms except the 5 of interest in the BM dataset.

Files created/downloaded:

|  |  |
| --- | --- |
| **Name** | **Description** |
| Data Entry - Bloods 29.11.23.xlsx | Updated version (I added the date) made by Sas on SharePoint that contains the second blood sample results. |
| Data Entry - Milk 29.11.23.xlsx | Updated milk data made by Sas on SharePoint that contains all the bulk milk sample results. |
| Data Entry – Questionnaire.xlsx | Raw data from SP. |
| Tracker - SeleniumFarms\_2023.xlsx | Spreadsheet to link questionnaire data (no supply number) with selenium data. I modified the tracker spreadsheet on SharePoint and manually added a column in the second sheet to capture the name used in the questionnaire. |
|  |  |
|  |  |

* Blood and bulk milk data are in separate spreadsheets.
* Blood dataset:
  + 651 observations of 11 variables from 60 SNs.
    - All unique.
    - Milk Se column is empty except 1 entry (removed column).
    - Notes covers off transcription errors and indicate 2nd test.
    - Unnamed columns 10 and 11 contained some vestigial notes from Sas’ data entry.
    - 3 farms have 2 supply numbers.
    - 50 2nd visit samples.
    - The 5 farms have 10 bloods from each visit.
  + Transcription error handling – see take 1.
  + Age is missing from 4 cows
    - 117 x 2 as above.
    - No tag on Farm H.
    - Unknown age on farm U.
  + All unique rows.
  + 60 farms (supply numbers), 7 regions.
  + 8, 10 or 11 bloods/farm, 3-11 farms/region.
    - 39655 had 8 bloods,
    - 45007, 45627, and 45632 had 11 bloods.
    - These are all legit and what the vets collected according to Sas.
  + Sampled 20 Sep to 04 Oct on V1, 7 Nov to 10 Nov on visit 2.
  + Age = 2-13, median = 4
  + Blood Se a little fat-tailed, mean = 1404, SD=613, range = 55-3253, some regional variation, large between-farm variation.
  + Se thresholds:
    - 7/601 (1%) of observations had blood Se < 250 nmol/L, all on 1 farm (45636 - Manawatu), which has the lowest mean blood Se.
    - 41/601 (7%) were below 500 on 12 farms.
    - 176/601 (30%) were below 1000 on 34 farms.
* BM dataset:
  + 181 observations of 3 variables after removing the 2 duplicates (take 1) (there were another 6 variables that I dropped as they were not germane according to Sas).
  + All unique.
  + 121 unique supply numbers – I understand a larger set of farms was sampled, so I restricted it to the 60 with blood data.
  + That meant there were 60 unique SNs, all with 2 visits.
  + I converted the one value of “<0.0025” to 0.
  + Only 50 unique values of Se, which are very low values without great precision.
  + Sampled 20 Sep – 3 Oct at V1 and 7 Nov – 13 Nov at visit 2.
  + BM Se is NOT ND (right tail), median=0.013, range = 0 -0.042
  + BM and serum dates generally disagree – up to 6 days out.

## Descriptive analysis

* 601 observations of 10 variables:

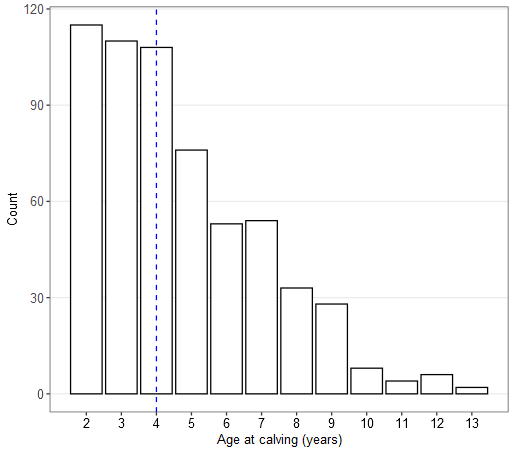
1. Supply number
2. Farm name
3. Region
4. Cow ID
5. Blood sample date
6. Cow age
7. Blood Se
8. Bulk milk test date
9. Bulk milk Se
10. Study ID

* Age is missing from 4 cows: 2 cows could not be differentiated from each other; 2 cows were missing age in raw data.
* All unique.
* 60 farms (supply numbers), 7 regions, 3-11 farms/region (table).

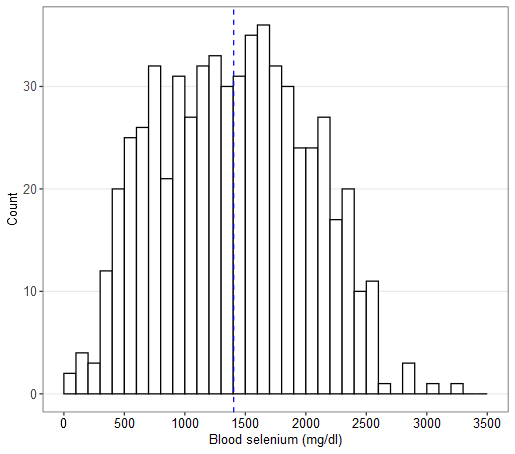
**A white paper with black text

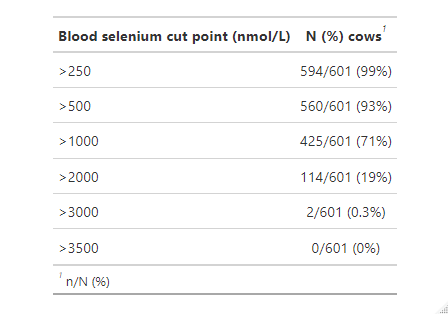
Description automatically generated**

* 1 farm had 8 bloods (Canterbury), 56 had 10, and 3 had 11 bloods (Manawatu).
* Sampled 20 Sep to 04 Oct.
* Age = 2-13, median = 4, not as many heifers as expected

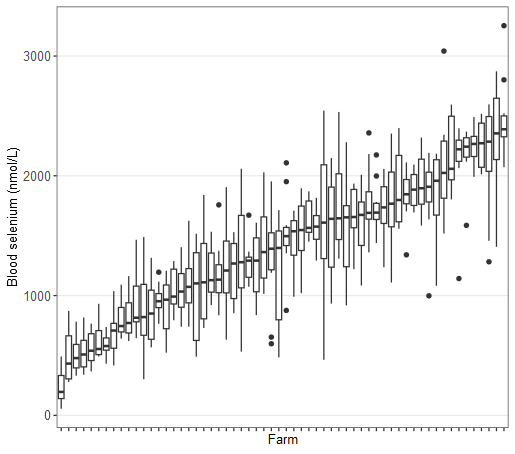


* + Blood Se ND, mean = 1343, SD=554, range = 55-2596, some regional variation, large between-farm variation.

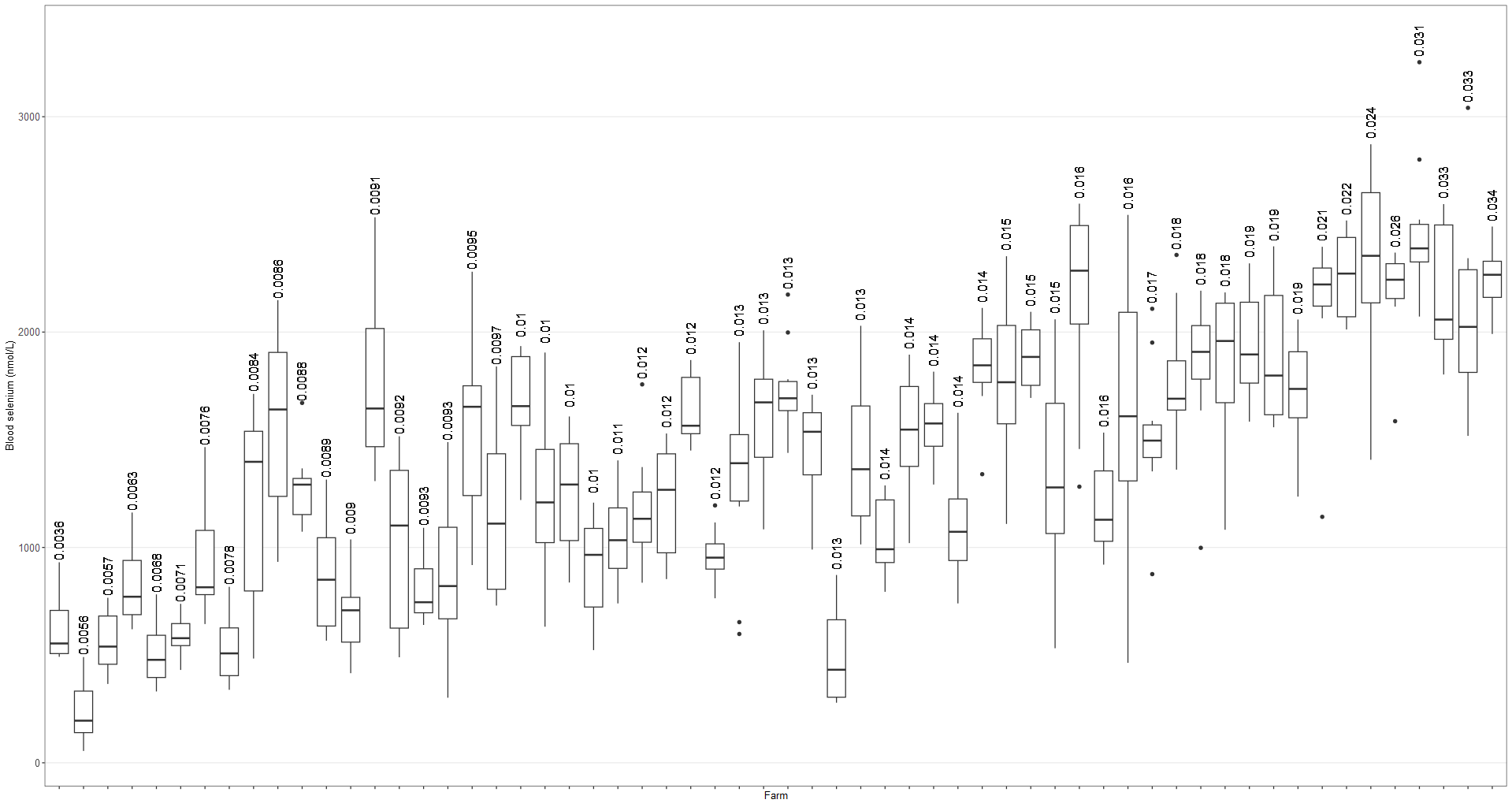




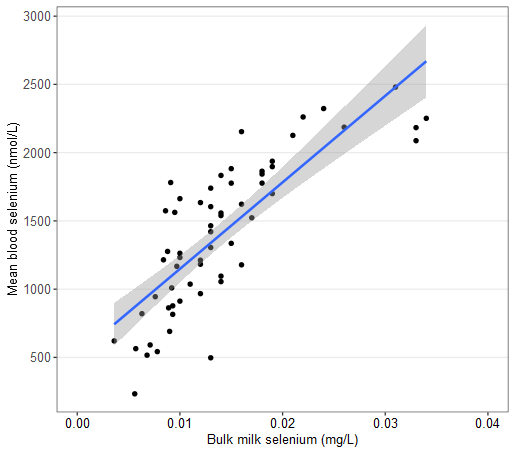
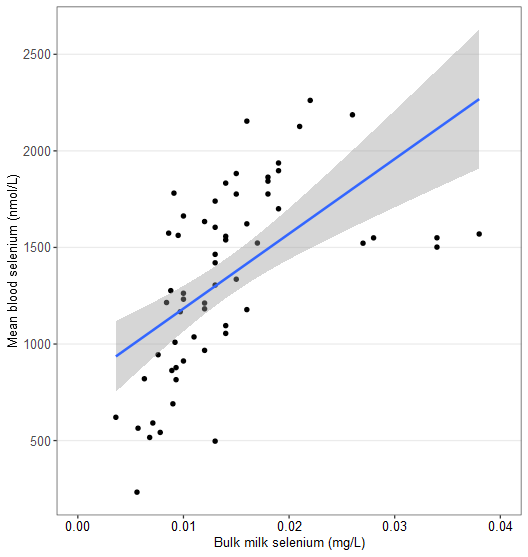
* Quite a bit of variation within and between farms in blood Se



* When sorted on bulk milk Se, there is clearly an overall positive association, but it shows the large within-farm variance in blood Se.



* With the updated dataset, when BM Se is plotted against mean blood SE, the positive relationship is clearer, but there are still 3 high (the highest) BM Se farms with lower-than-expected blood Se:
  + 36765 (Farm D, Canterbury),
  + 36843 (Farm G, Canterbury),
  + 42594 (Andrew Moffitt, Taranaki).
* Plot on left = old, plot on right = updated.
* Looks like a curvilinear relationship.

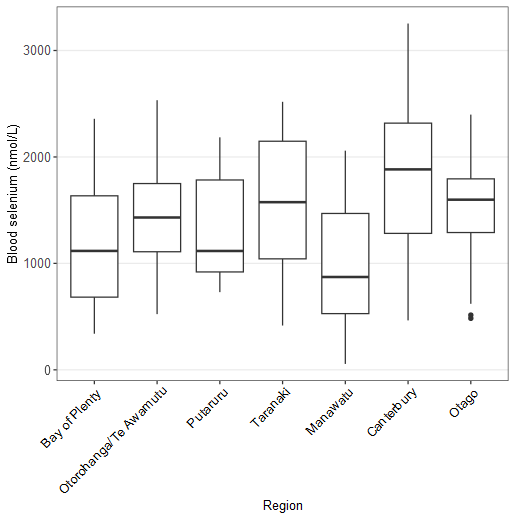


* No clear association with age:

A graph of a number of lines

Description automatically generated with medium confidence

* No clear association with region, though Manawatu was lower:

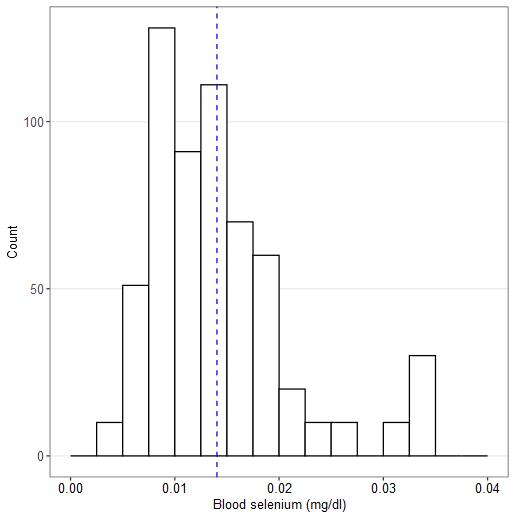


* Age doesn’t seem to be associated with region:

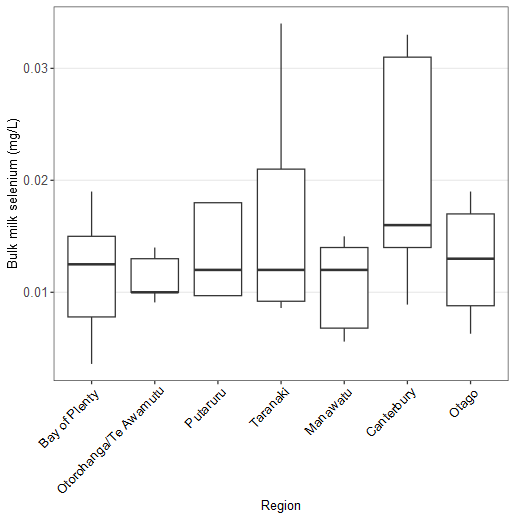
A graph with lines and dots

Description automatically generated with medium confidence

* BM dataset:
  + BM Se is NOT ND (right tail), median=0.013, range = 0.0036-0.034

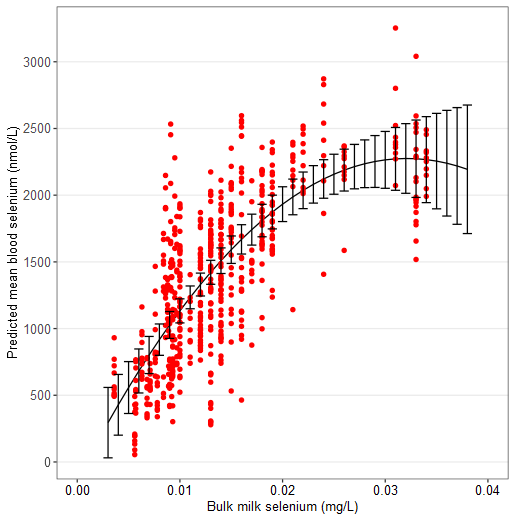


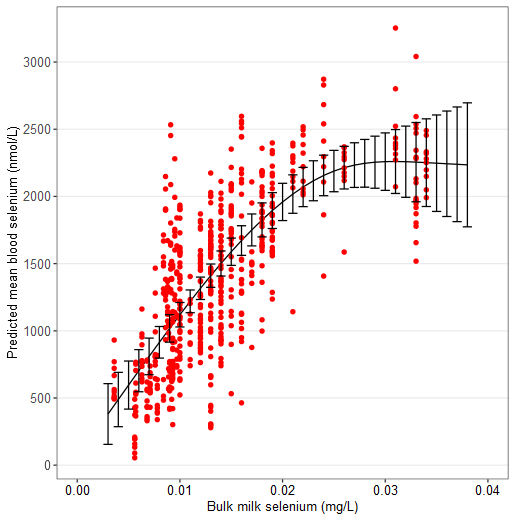
* + BM and serum dates generally disagree – up to 6 days out.
  + No clear association between BM Se and region: though Canterbury seems high and Manawatu low:



## Association between serum and BM Se

* I scaled BM Se by multiplying it by 1000 (converting mg to g) to show the effect of a 1mg increase in BM Se instead of a 1g increase, which makes no sense.
* Simple linear models, no other variables because I am taking mean blood Se, so age is irrelevant.
* BM Se led to a non-linear model. Adding a square term improved fit (LR test and AIC) and resolved nonlinearity. Adding a cubed term made no improvement.
* Adding a spline at 0.025 was slightly better than a polynomial, but the AICs were 859.0963 (polynomial) and 858.8876 (spline) – hardly a difference. I find splines hard to interpret.
* Diagnostics
  + Residuals are ND, homoscedastic, and linear.
  + Farms 45632 (Manawatu) and 74590 (Oto/TA) were most influential – low/high blood Se given BM Se respectively.
  + Farms 42594 and 36765 have large leverage due to highest BM Se.
  + Farm 45632 had the highest residual (lower blood Se than expected).
* Prediction – plots of polynomial model (L) and spline model (R):





# Questionnaire

* Questionnaire is in 2 sections:
  + “General farm information” (9 questions),
  + “Supplementation, signs and testing” (21 questions).
* Data
  + Milk production (1.6): took midpoint of range, multiplied cow-level values by N cows at peak, values expressed per day were set as NA.
  + NICR (1.8): midpoint of ranges taken.
  + SCC: midpoint of ranges taken.
  + Supplementation (2.1): 4/5 of the “Sometimes – depends” answers were in fact “Yes” – change to Yes. The one “Sometimes” farmer stated “Maybe in mineral mix”.
  + 60 observations, 85 questions.
  + Breed (Q1.5) categorised into F ("F10", "Fr", "Fr - FX", "HF"), J, and XB ("Kiwi X", "FX", "Jersey X", "Jersey, Kiwi X", "Kiwi X", "Mixed", "X Breed", "XB-Fr", "Fr, Kiwi X").
  + Milk production (Q1.6): took midpoint of ranges when applicable, multiplied cow-level answers (e.g., “330/cow”) by N cows at peak, values expressed per day were set as NA.
  + NICR (Q1.8): midpoint of ranges taken.
  + SCC (Q1.9): midpoint of ranges taken.
  + I made a table of general farm information.
  + I made a graph function for plotting responses to questions with an “other” category.
  + I made a function for computing median (IQR) count of options each respondent selected to a question.
* Q2.1:
  + 4/5 of the “Sometimes – depends” answers were in fact “Yes” – change to Yes.
  + The one “Sometimes” farmer stated “Maybe in mineral mix”.
* Q2.2:
  + 8 options, median of 2 (range = 1-6) selected.
  + Injectable most common, 23/43 used B12/Se (but 1 was only in calves), 5 Selovin, 3 Multimin, 4 not specified (e.g., “Pre-calving”), 1 Vijec.
  + 1 “other” = lick blocks
* Q2.3 (how much Se):
  + Free text only.
  + ChatGPT.
* Q2.4 (Why choose method):
  + 4 options, 1 skipped, median of 1 (0-4) selected.
  + “Other” options = bloods more even (1), farm consultant (1), vet advice (2), easy (2), trying something different (1), keeps water clear (1).
* Q2.5 (cost):
  + ChatGPT.
* Q2.6 (have you changed method):
  + None skipped.
  + 68% no.
  + 2 people with “no” gave reason: “Nothing else has changed significantly” and “Think it is effective”.
  + All bar 2 with “yes” gave reason: produce table as answers are mostly separate.
* Q2.7 (If you use multiple methods of supplementation for your dairy cattle - why do you do that?):
  + Only those who used >1 method responded.
  + ChatGPT.
* Q2.8 (What is the main reason you choose to supplement your dairy cattle?):
  + 1 skipped but wrote “Fert reps advice” in details for “other, so I will change it to “other”.
  + Median of 2 options chosen (1-5).
  + Provide “other” options in a table.
* Q2.9 (Have you ever seen any signs of selenium deficiency in your dairy cattle?):
  + 0 skipped.
* Q2.10 (Have you ever seen signs of selenium deficiency in your neighbours stock?):
  + Only 1 answer possible.
  + 10 skipped.
  + 1 said too busy to look at neighbour or stock = convert to “NA”, meaning 11 skipped.
* Q2.11 (What do you think are the signs of selenium deficiency in cattle?):
  + 2 skipped.
  + ChatGPT.
* Q2.12 (What do you think the impacts are of lower that optimal selenium in your herd?):
  + 5 skipped.
  + ChatGPT.
* Q2.13 (If you have had selenium deficiency in your dairy cattle, how did you know?):
  + Not applicable to 30 (50%).
  + Of those who have had deficiencies, 0 skipped, and a median of 1 (1-2) responses were given.
  + 3 entered an answer under “other”:
    - Never tested (1),
    - “I don't think I have as I have always supplemented it” (1),
    - “White muscle disease. Bulls couldn't really move, weak.” (1)
* Q2.14 (Have you ever tested your dairy cattle for selenium concentration in the liver or blood?):
  + 4 options; could choose more than one.
  + 0 skipped, median of 1 response (1-3).
  + 12/60 selected “no” and 3/60 selected “unsure”. Therefore 75% had done either blood or liver testing.
* Q2.15 (If you have tested, what values do you think are ideal? (If you are not sure, just circle unsure)):
  + None skipped, 15 entered not applicable, median of 1 answer (1-2).
* Q2.16 (When was the last time you tested your dairy cattle?):
  + 9 skipped.
* Q2.17 (Why do you choose not to test your dairy cattle for selenium? Or what are your reasons for a reduced frequency of testing?):
  + I edited the Q17 question because the questionnaire actually used didn't contain "Or what are your reasons for a reduced frequency of testing?".
  + Changed one response that had not selected any options but provided a comment to “other”.
  + 37 not applicable, 23 answered.
  + 0 skipped, median 1 option (1-3).
  + Present “other” responses as a table as there are only 9.
  + Of the 11 who wrote a comment, 3 chose a non-other option and 8 chose “other”.
  + Of the 23 who responded, 9 did not respond to Q16. The other 14 had a mixture of responses to Q16 (6 = 1-3 years ago, 5 = >3 years ago, 3 = within last year).
  + Of the 9 who did not respond to Q16, all responded to Q17.
* Q2.18 (How do you know if your supplementation program is working?):
  + 1 did not respond (NAs – new herd – Matt Mofitt, Taranaki) because it is a new herd and has not had a chance yet.
  + Otherwise, none skipped, median of 1 response (1-2).
  + 2 gave an “other” response.
  + When signs of deficiency was selected (16), the explanations were:
    - None (6),
    - Repro performance (5),
    - Clean post calving/RFM/sick cows (3),
    - By testing (1),
    - Low empty rate + not many sick cows (1).
  + For “other”, comments were:
    - Used to test years ago (1).
    - No response (1).
* Q2.19 (Do you test your pasture or soil for selenium?):
  + 0 skipped, 1 (1-2) responses.
* Q2.20 (Why do you choose not to test your pasture or soil for selenium?):
  + 0 skipped, 41 not applicable, 1 (1-2) responses.
  + Of the 16 who answered “no” to Q2.19, 15 gave an explanation and 1 did not (no reason).
  + Of those who did not answer “no”, 2 answered “Too expensive” but selected yes- soil, 1 answered “Too expensive” and “other” but selected yes – soil, 1 answered “other” (“Not aware of the possibility as it is not in the standard testing procedures”) but selected “unsure”
  + Focusing on only the 15 who answered “no” to Q2.19 and responded to Q2.20, 2 selected 2 responses, so there are 17 responses. 9 (60%) selected “Don't see the need - I supplement anyway”, 4 (26.7%) selected “other”, 3 (20.0%) selected “I would, but I don't get around to it and then I forget”, 1 (6.7%) selected “Don’t see value in the information”, and none selected “Too expensive”. Among the “other” responses, 2 hadn’t/never thought about it, 1 was “Too lazy” and 1 said “Bloods are reliable”.
* Q2.21 (What/who is the main source of information for you to get advice about minerals and trace elements?):
  + 1 skipped but should have been “other” as there was a comment, so I corrected it.
  + 1 (1-3) options chosen.
  + All those who chose “other” (15) provided an explanation:
    - Feed agents/reps/companies/fertiliser reps (11),
    - DairyNZ (1),
    - Independent soil/pasture representative (1),
    - Nutritionist (2)

# Associations between the questionnaire and selenium concentrations

* Questions that would make sense to include:
  + 2.1 (do you supplement): 58/60 answered “yes” = no point analysing it.
  + 2.2 (how do you supplement): more than one option could be picked = challenging to work out what drives any differences we find.
  + 2.14 (have you ever tested your cattle): same as 2.2 (more than one option).
  + 2.16 (when was the last time you tested cattle): OK but 69% within last year.
  + 2.19 (do you test pasture/soil): same as 2.2.
  + 2.20 (why choose not to test pasture/oil): same as 2.2.
* Questions not included:
  + 2.1 (all bar one respondent answered “yes”).
* Questionnaire dataset does not have supply numbers! I added them manually via the “Tracker” spreadsheet.
* Make boxplots of blood and BM Se for each option of each question.

# References