DAIRYLAND
Laboratories, Inc.

On Validation: Verifying Accuracy and Validity of NIRS-predicted Results

Phil Goldblatt (NIR Manager – Dairyland Laboratories)

Why Do We Test Forages with NIRS Models?

- Making a ration to optimize animal performance
- Timing harvest
- Pricing for hay
- Plant Breeding

… To quantify variation in order that we may <u>make a decision or draw a conclusion</u>.

How Do We Know Our NIR Models Help Us Make The Right Decision?

"…accuracy of NIR predictions cannot be assumed because a "number is produced" by the NIR solution."

"Conducting an independent ***validation*—comparing laboratory-measured to NIR-predicted values—is the ultimate standard for assessing the accuracy and reliability of NIR solutions."**

Castillo, M. S., Griggs, T. C., Digman, M. F., Vendramini, J. M. B., Dubeux, J. C. B., & Pedreira, C. G. S. (2025). Reporting forage nutritive value using near-infrared reflectance spectroscopy. *Crop Science*, *65*(3), 1–2. https://doi.org/10.1002/csc2.70063

What I'll Talk About:

1. What is "Validation?"
    1. "Calibration statistics," and why they aren't good enough
    2. Define and show validation
2. How do we (people who make and provide calibrations) do validation?
    1. Show examples of how validation is done in practice
3. Do I need to validate more?
    1. Is the validation done by your calibration provider enough for your use case?

What I'll Try to Not Talk About:

1. Too Much Statistics
2. Too Much Calibration/Model Development

DAIRYLAND
Laboratories, Inc.

What is Validation?

DAIRYLAND
Laboratories, Inc.

What Are “Calibration Statistics?”

Edit - P:\app\NIRS2\plcor13.eqa

File Name: P:\app\NIRS2\plcor13.eqa Equation File File Date: Thu May 23 12:07:34 2013 Last Upda
Serial No: 21250432 Constituents: 28 Calculated Equations: 0 Number of Variable
Segment 1 1100 - 2498 2

Constituent	Type	N	Mean	SD	Est. Min	Est. Max	SEC	RSQ	SECV
PROTEIN	1	4063	8.7464	1.4702	4.3358	13.1571	0.3628	0.9391	0.3673
ADF	1	735	5.3550	4.0648	0.0000	17.5494	0.7327	0.9675	0.7648
NDF	1	485	13.1551	8.2265	0.0000	37.8347	1.5121	0.9662	1.6298
ADP	1	745	0.6980	0.3629	0.0000	1.7867	0.2858	0.3799	0.2890
CA	1	1740	0.0416	0.0266	0.0000	0.1213	0.0190	0.4903	0.0192
P	1	1730	0.3227	0.0803	0.0817	0.5636	0.0370	0.7878	0.0380
MG	1	1744	0.1389	0.0364	0.0298	0.2481	0.0181	0.7536	0.0184
K	1	1741	0.4730	0.1139	0.1313	0.8148	0.0579	0.7415	0.0591
INSOL	1	1488	5.5853	1.9693	0.0000	11.4933	0.6499	0.8911	0.6662
ASH	1	1869	1.6547	0.5451	0.0193	3.2901	0.2602	0.7721	0.2674
FAT	1	1612	4.4389	1.7029	0.0000	9.5478	0.4463	0.9313	0.4507
pH	1	1235	4.1524	0.2228	3.4839	4.8209	0.0854	0.8531	0.0898
SUGAR	1	589	2.8254	1.8475	0.0000	8.3681	0.8308	0.7978	0.8767

Statistics about the population used to make the prediction model.

Product file description:

Parameter	PROTEIN	ADF	ADP	CA
Number of samples to Keep	147	150	149	152
Number of samples Rejected	16	13	14	11
Number of factors	5	6	7	5
Minimum of reference data	8.991	9.292	0.797	0.582
Mean of reference data	15.895	22.988	1.174	1.159
Maximum of reference data	20.650	40.101	1.587	1.777
SD of reference data	2.171	5.693	0.157	0.214
SEC	0.403	0.992	0.082	0.086
RSQ	0.966	0.970	0.729	0.836
SECV	0.450	1.126	0.096	0.101
RSQV	0.940	0.948	0.606	0.739
GDAve	1.000	1.000	1.000	1.000
NDAve	0.172	0.224	0.233	0.172
ExpNND	0.366	0.503	0.519	0.398
MinSED	0.000	0.000	0.000	0.000
Gain/Loss	0.000	0.000	0.000	0.000
SNV				

… or statistics about how well the prediction model explains the variation in the population used to make the model.

DAIRYLAND
Laboratories, Inc.

… And Why Are They of Limited Use?

- Does the alfalfa NDF calibration work well on western hay?
- Does this cereal silage calibration work on boot stage wheat?
- Does this corn silage calibration work on samples from Chile?
- Can you use your manure calibration on horse manure?
- Can this mixed hay calibration handle a mixture of oats, kochia, and hemp?
- Does this corn silage calibration work with this year's newest hybrids?

Calibration statistics answer none of these questions!

(They only answer questions about statistics related to the population used)

DAIRYLAND
Laboratories, Inc.

What is “Validation?”

“The use of NIR spectroscopy to estimate nutritive value and digestibility of forages can be traced in US literature to the seminal study by Norris et al. (1976). Recognizing the importance of validation for NIR-predicted values, Norris et al. (1976) indicated that “to provide a better test of the merits of such measurement (referring to NIR-predicted values), we used the odd-numbered samples to develop calibration equations, which were then used to predict crude protein (CP), neutral detergent fiber (NDF), and in vivo dry matter digestibility (DMD) for the even-numbered samples.””

Castillo, M. S., Griggs, T. C., Digman, M. F., Vendramini, J. M. B., Dubeux, J. C. B., & Pedreira, C. G. S. (2025). Reporting forage nutritive value using near-infrared reflectance spectroscopy. *Crop Science*, *65*(3), 1–2. https://doi.org/10.1002/csc2.70063

DAIRYLAND
Laboratories, Inc.

What Exactly Did Karl Norris Do Here?

Norris took **spectra of samples not in the models** and used the model to make NIR predictions. He then compared those predicted values to the **reference chemistry (same methodology) values for those samples.**

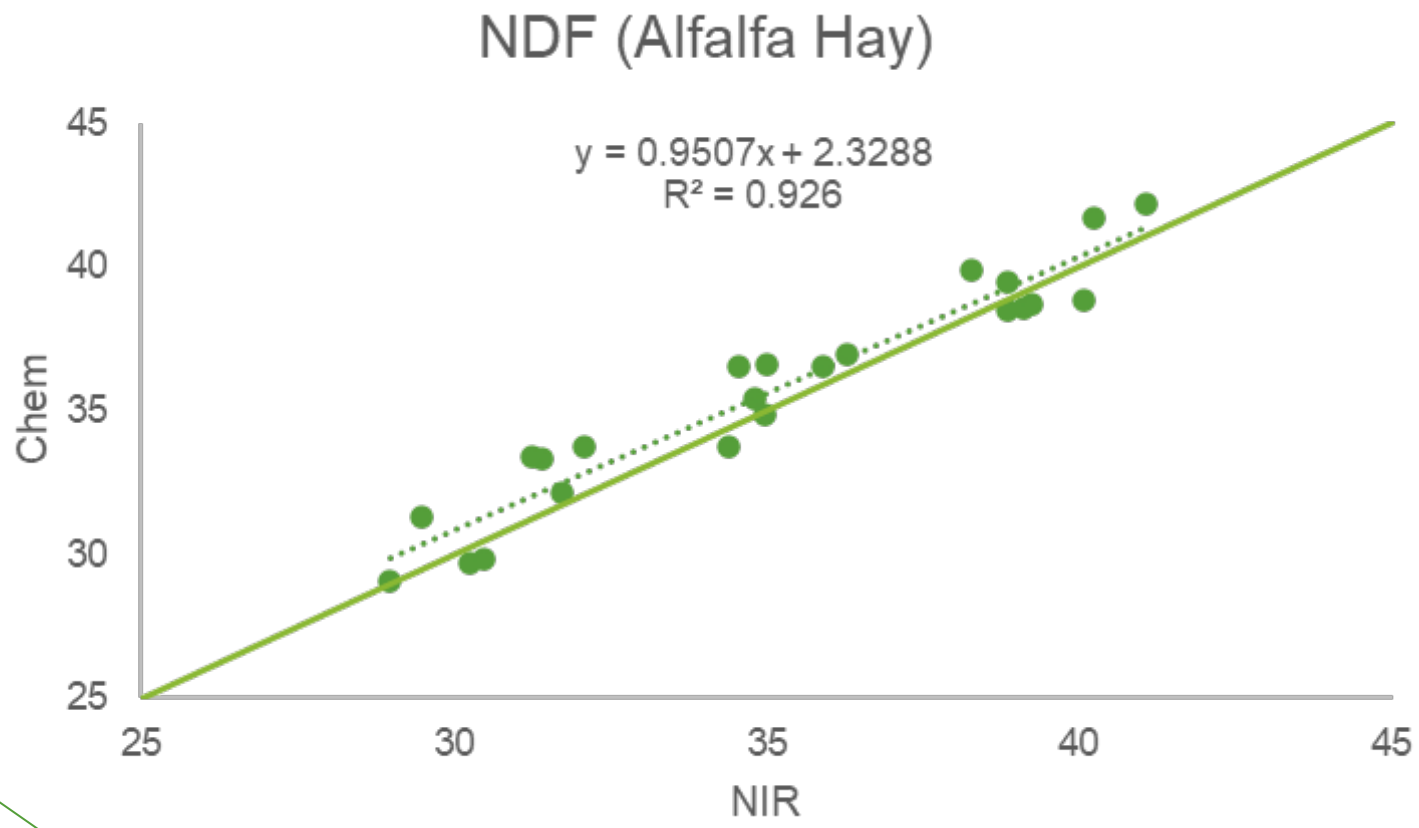
"First, the samples for the independent validation exercise were not included in the calibration of the NIR model and, second, the samples represent the population being analyzed and cover the range of NIR-predicted values."

Castillo, M. S., Griggs, T. C., Digman, M. F., Vendramini, J. M. B., Dubeux, J. C. B., & Pedreira, C. G. S. (2025). Reporting forage nutritive value using near-infrared reflectance spectroscopy. *Crop Science*, *65*(3), 1–2. https://doi.org/10.1002/csc2.70063

DAIRYLAND
Laboratories, Inc.

A Validation Set…

	Chem	NIR	Chem-NIR	(Chem-NIR)^2
130024	29.66	30.26	-0.60	0.36
123129	42.20	41.08	1.12	1.26
113995	35.45	34.81	0.64	0.41
146577	38.83	40.06	-1.23	1.50
123146	41.73	40.24	1.49	2.21
113910	34.84	34.98	-0.14	0.02
155801	33.75	32.08	1.67	2.78
155802	33.34	31.41	1.93	3.71
130092	33.42	31.24	2.18	4.74
155142	38.60	39.10	-0.50	0.25



Square root of average of these to get RMSE (1.20)

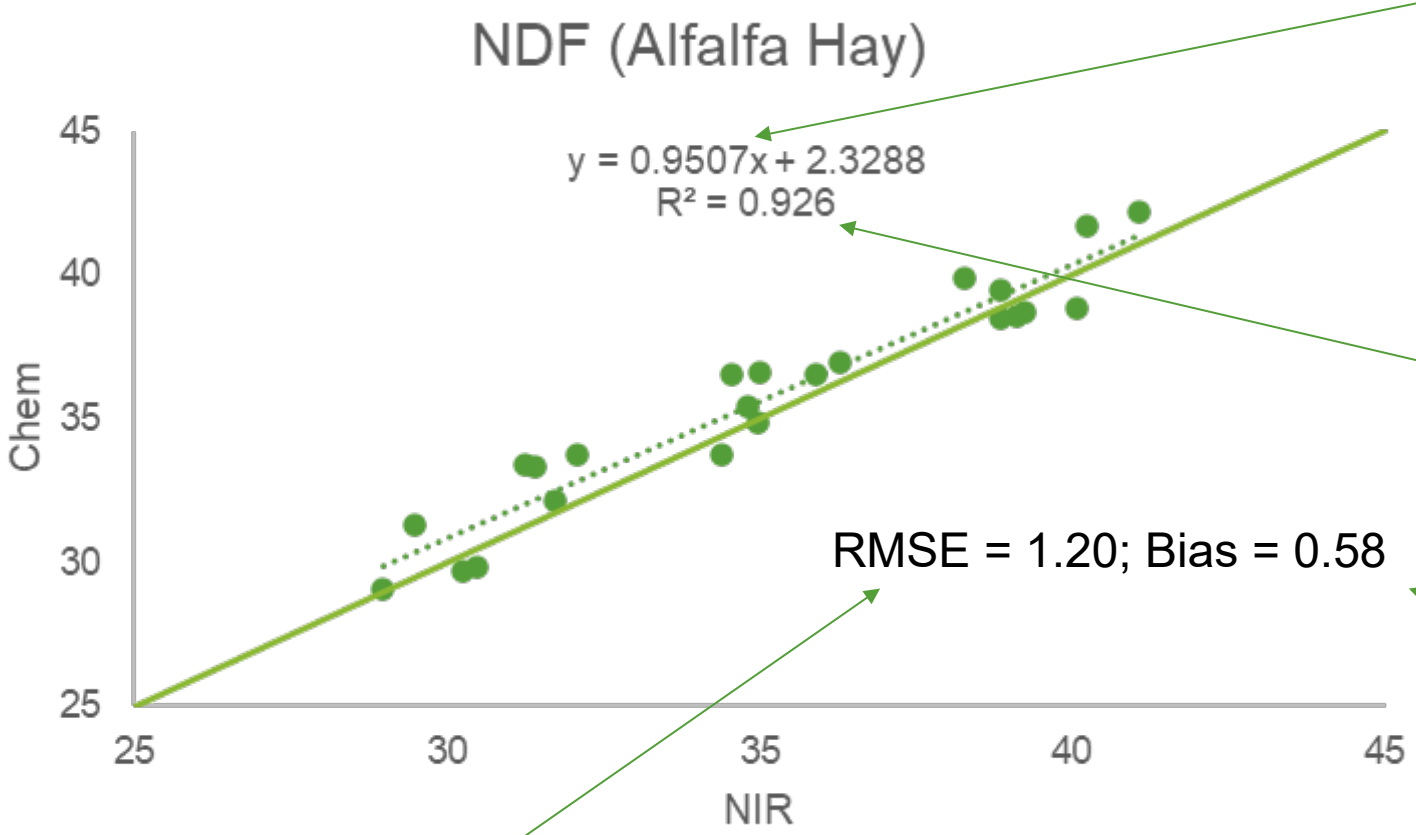
Reference Chemistry
(Same method as calibration)

Average these to get a BIAS (0.58)

NIR-Predicted Values

DAIRYLAND
Laboratories, Inc.

A Validation Set…



Slope: Want it close to 1, which means that there is no systematic error with regard to magnitude of reference value.

R^2: Closer to 1 is better, shows correlation between Chemistry and NIR – dependent on range of values.

RMSE = 1.20; Bias = 0.58

RMSE: Root Mean Square Error - should be slightly higher than error of reference method for good model. You can expect about 2/3 of samples to be within this of the reference value.

Bias: On average, are we high or low?

DAIRYLAND
Laboratories, Inc.

We Must *Confirm* the Number Produced by NIR is Accurate (through validation):

- Scan samples not in model (preferably samples in your decision unit) and record predicted values
- Get reference values from same reference method used to make model
- Calculate error, bias, slope, and r^2

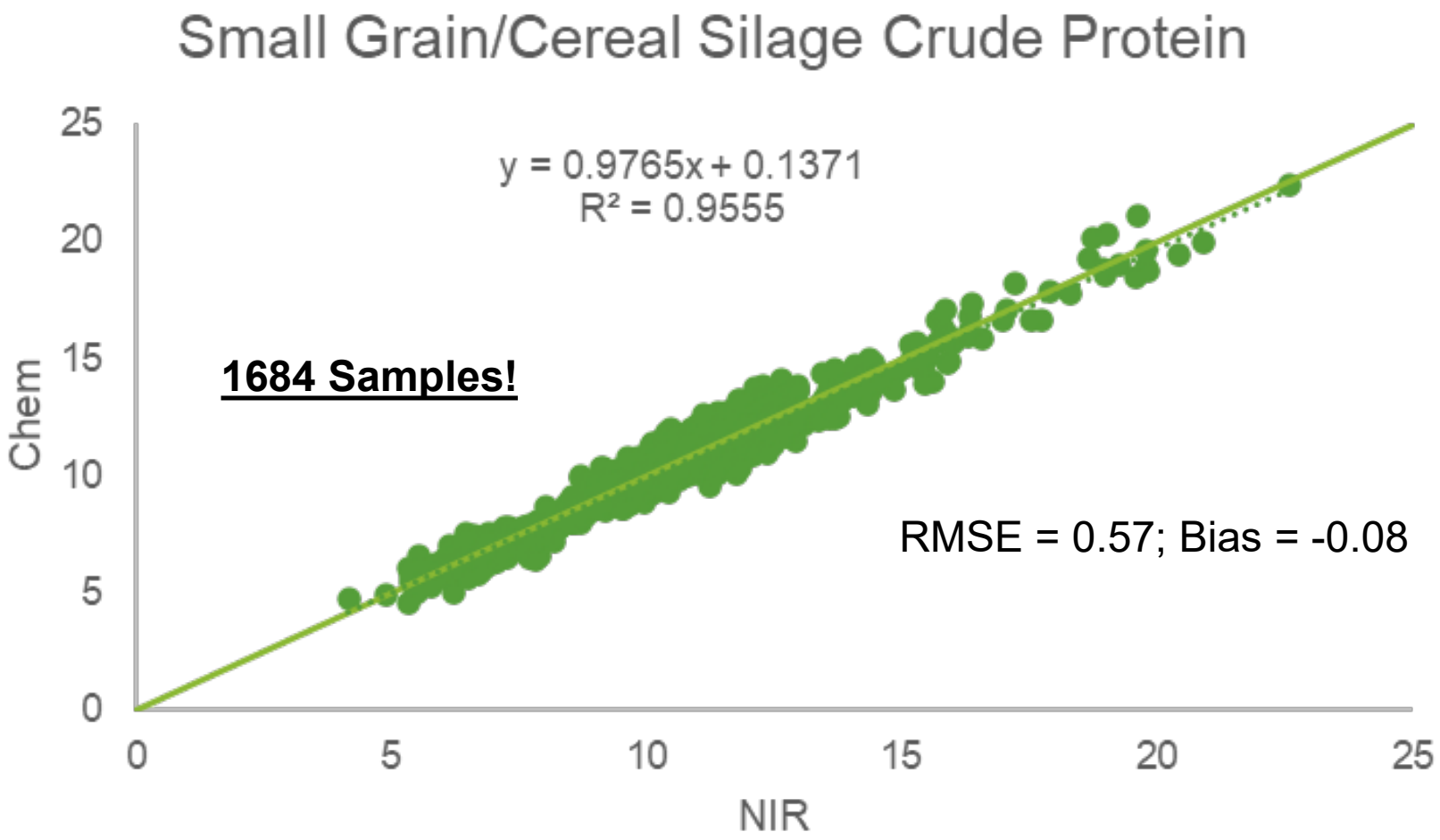
"Do ***I*** have to do this for every calibration I have!?"

DAIRYLAND
Laboratories, Inc.

How Do Calibration Providers Validate?

DAIRYLAND
Laboratories, Inc.

Your Calibration Provider Has (Probably) Validated!



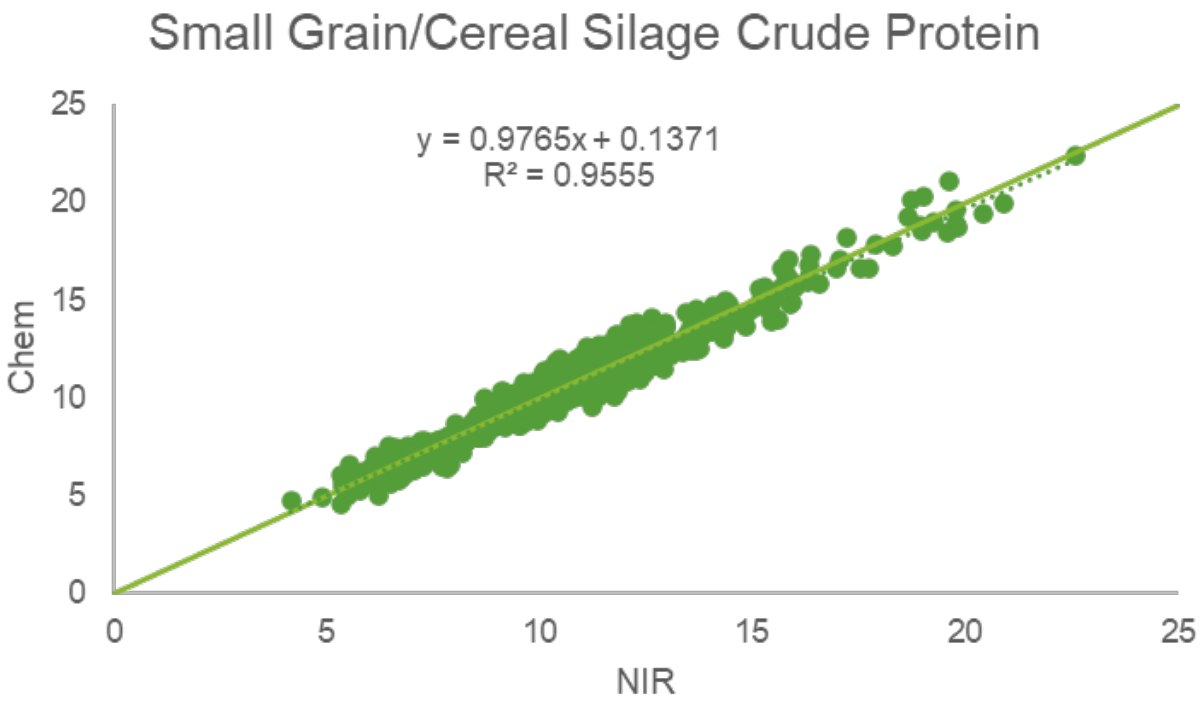
In practice, a calibration provider will have validated already, using all available samples with reference chemistry (or a sufficiently large set).

Providers will use validation sets heavily to optimize performance and select the best model to release.

The calibration provider will validate on EVERYTHING, not just your samples!

DAIRYLAND
Laboratories, Inc.

Is Performance Consistent Across Subpopulations?



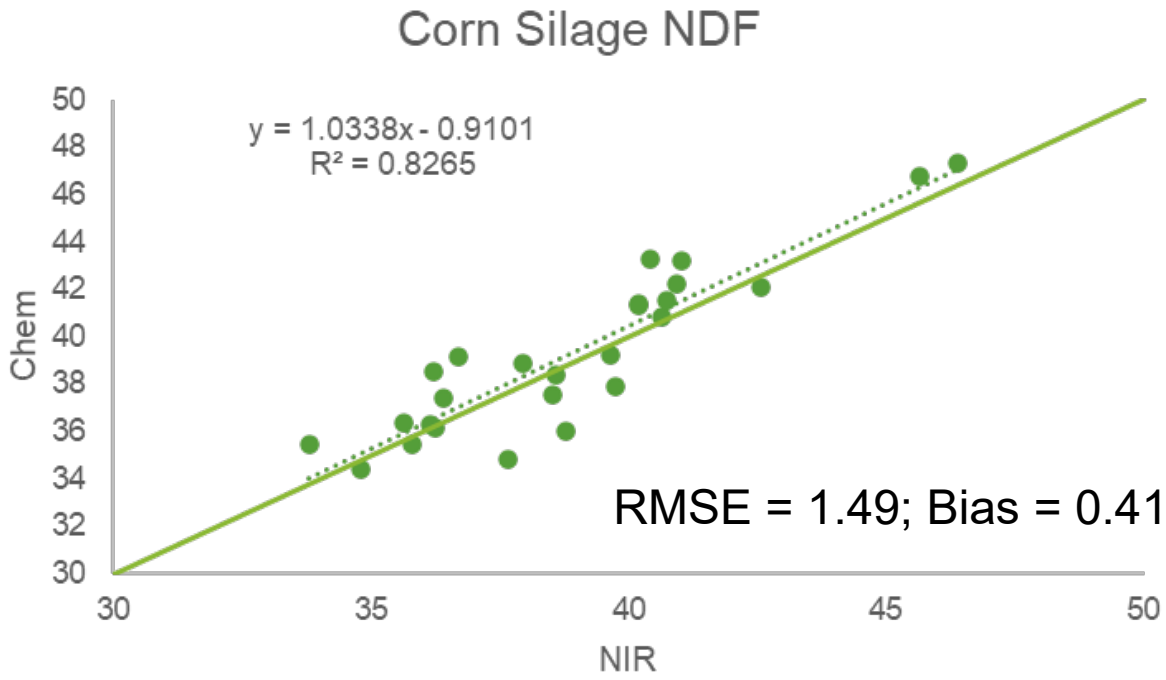
Subtype	RMSE	BIAS	Slope	r^2
Wheat	0.55	-0.03	0.98	0.95
Oat	0.68	-0.08	0.98	0.92
Barley	0.46	0.04	1.13	0.96
Sorghum/Sudan	0.63	-0.14	1.01	0.94
Millet	0.62	0	0.96	0.91
Triticale	0.64	-0.16	1.09	0.94
Rye	0.53	0.02	0.93	0.9
Triticale/Pea	0.57	-0.04	0.94	0.95
Unknown	0.58	-0.05	0.95	0.98

Different similar species may be lumped together into one calibration for the sake of practicality or out of necessity – a provider should verify that performance is consistent across types.

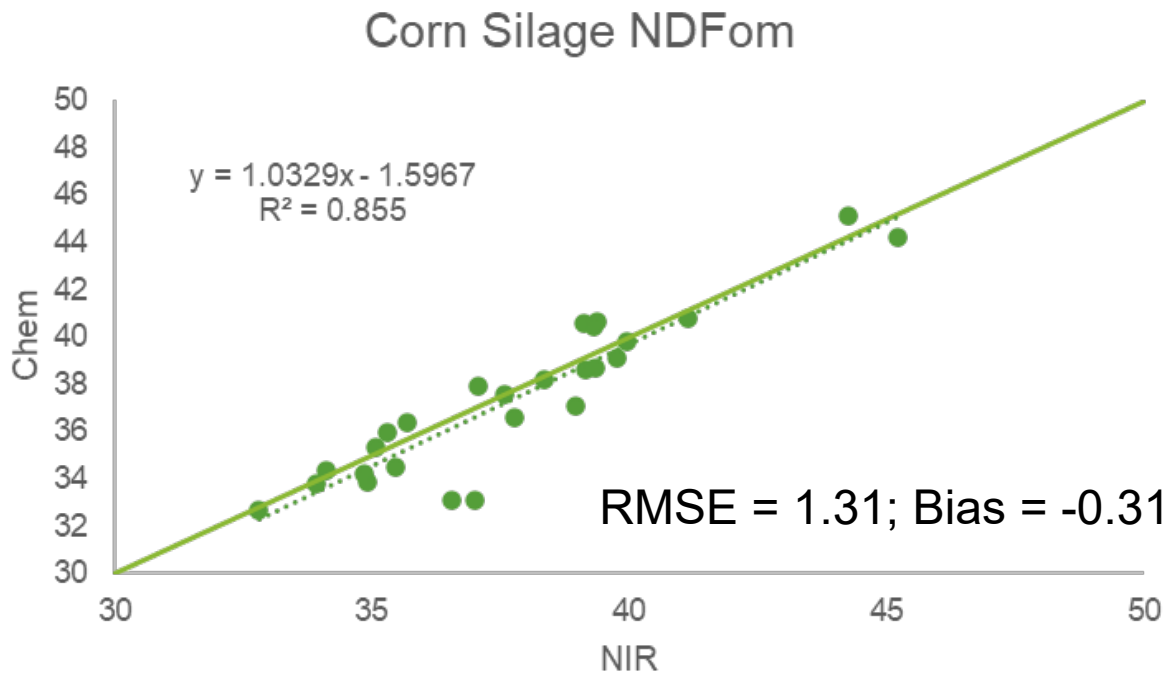
Other subpopulations include: samples from different regions, different instruments, labs, etc.

DAIRYLAND
Laboratories, Inc.

Do Results Make Sense In Context With Other Nutrients?



RMSE = 1.49; Bias = 0.41

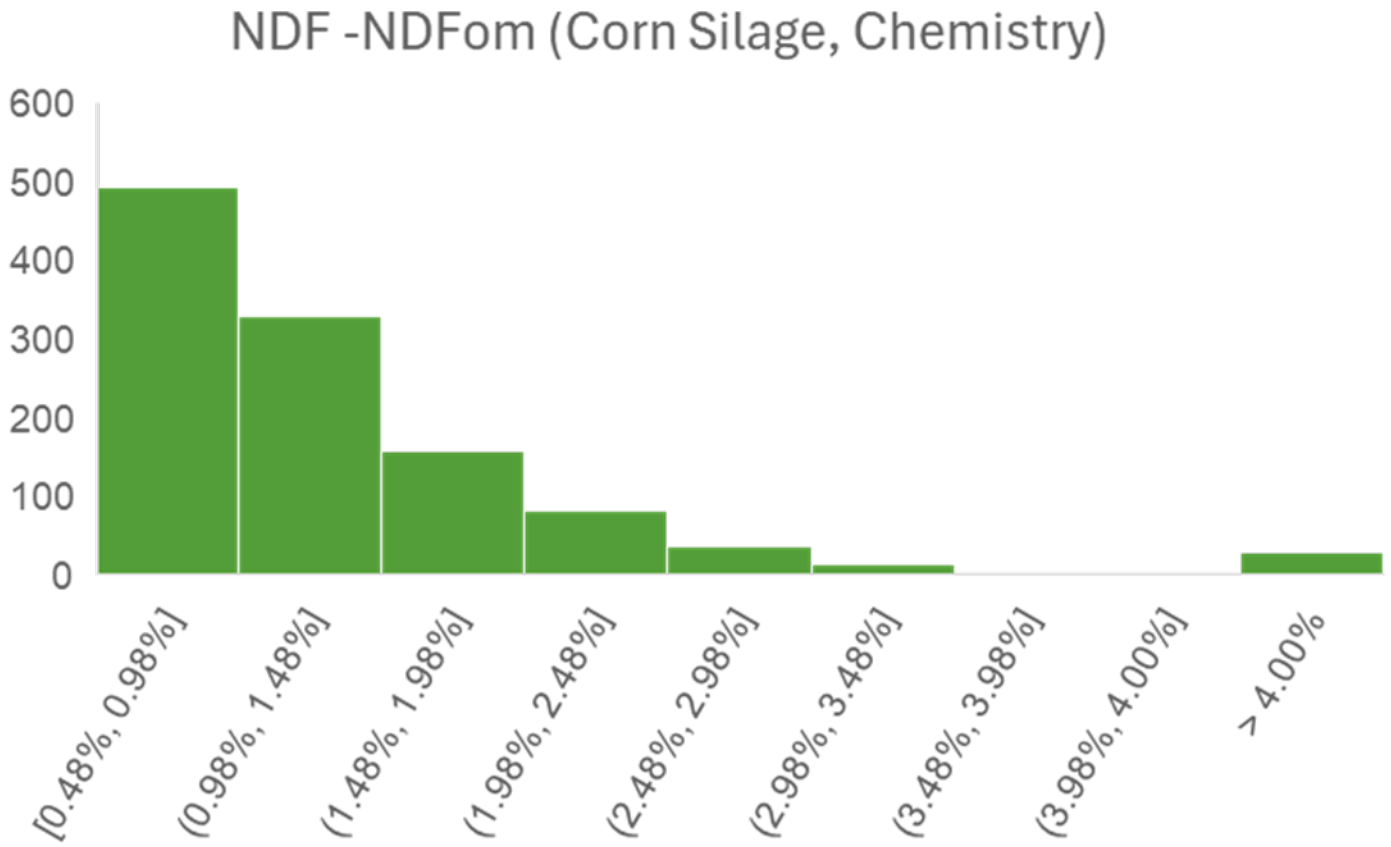


RMSE = 1.31; Bias = -0.31

Validating NDF doesn't just mean making sure that NDF is accurate… NDF must be greater than NDFom or it is "wrong."

DAIRYLAND
Laboratories, Inc.

Nutrients May Be Close, Have Required Order



Average = 1.31%

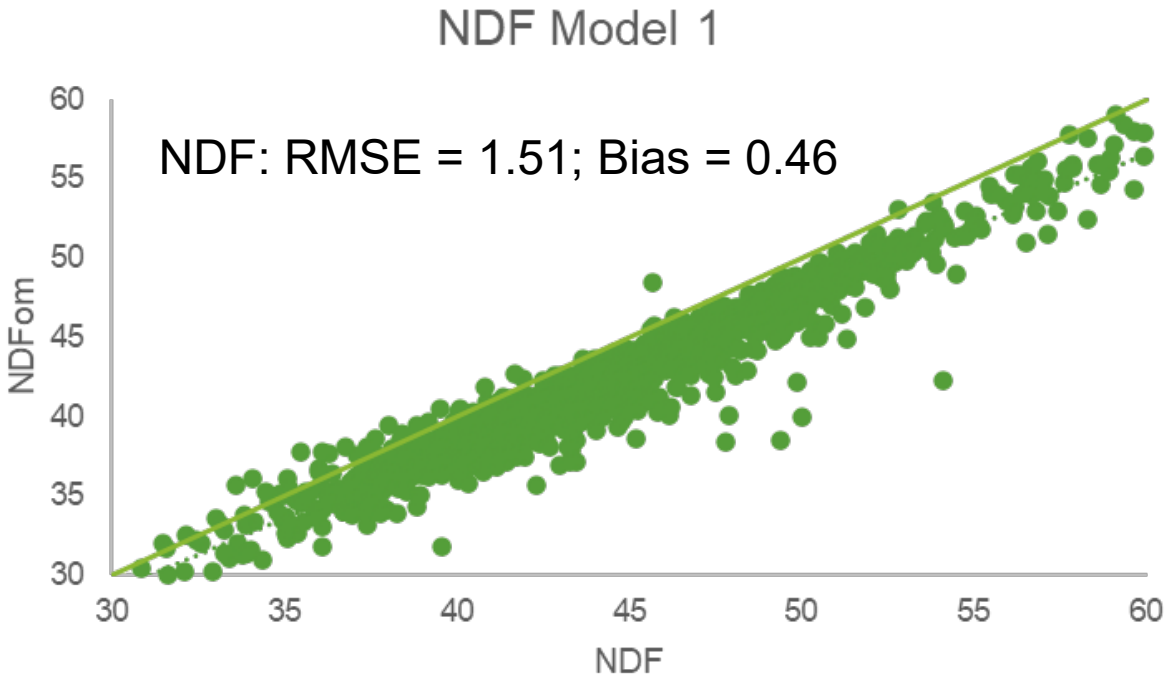
Average difference (by reference chemistry method) between NDF and NDFom is smaller than NDF RMSE…

Random variation can produce "out-of-spec" results even if raw accuracy/bias is acceptable.

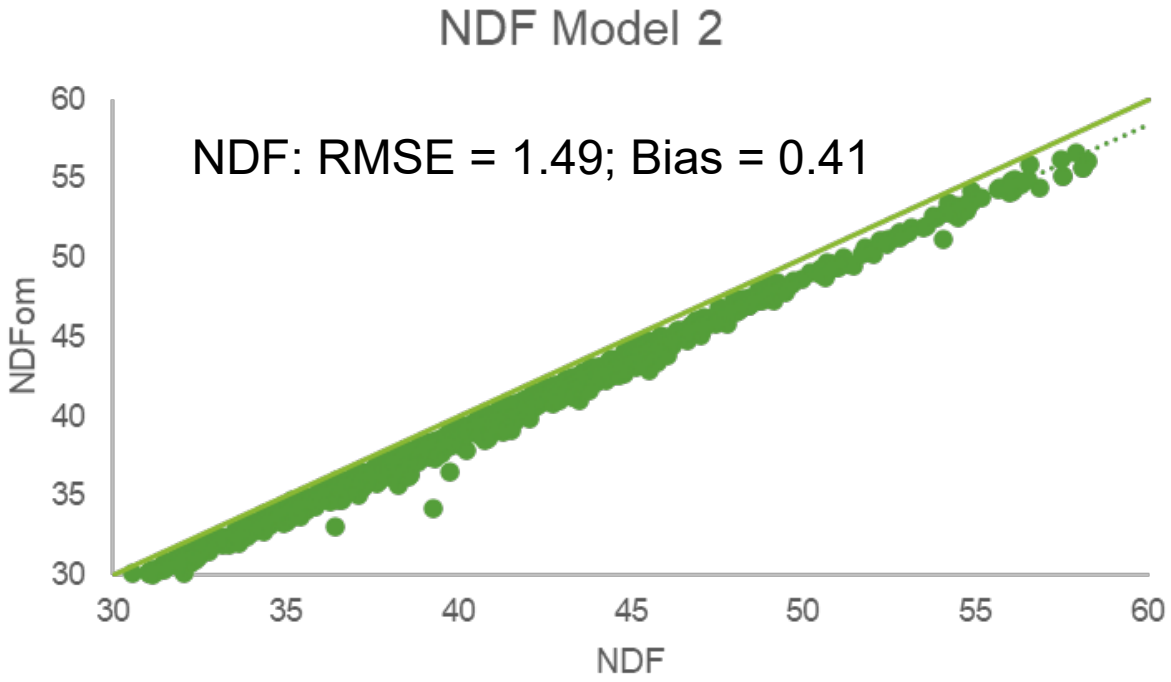
Other examples: uNDFom30/48, ADF/NDF, CP/Amino acids, etc.

DAIRYLAND
Laboratories, Inc.

Preserving Nutrient Relationships Can Be Tough



3.17% of Samples Out-of-Spec



No Samples Out-of-Spec!

When creating/updating models, making predictions with large numbers of spectra from a subpopulation to scrutinize relationships is wise – no chemistry needed!
Error/bias are similar with both models, but model 2 is a big improvement.

DAIRYLAND
Laboratories, Inc.

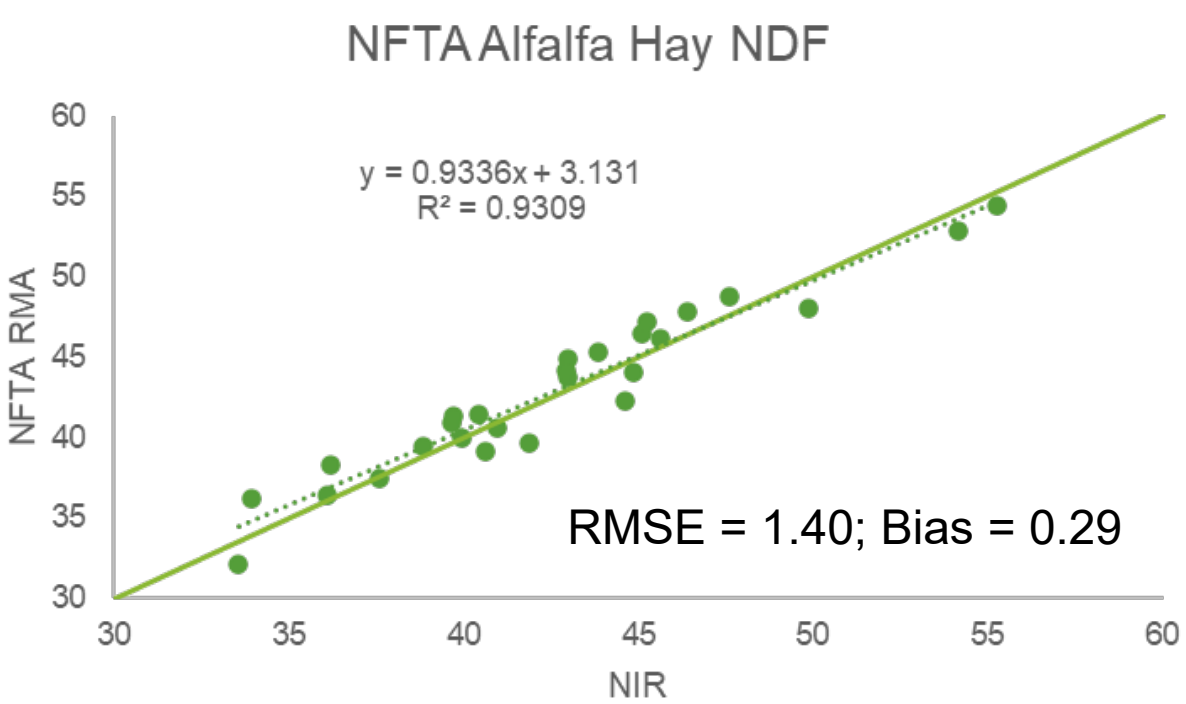
“Does The Calibration Work on My Samples?”

May be asked when provider enters a new lab/geography - In reality, this is a 3-part question:

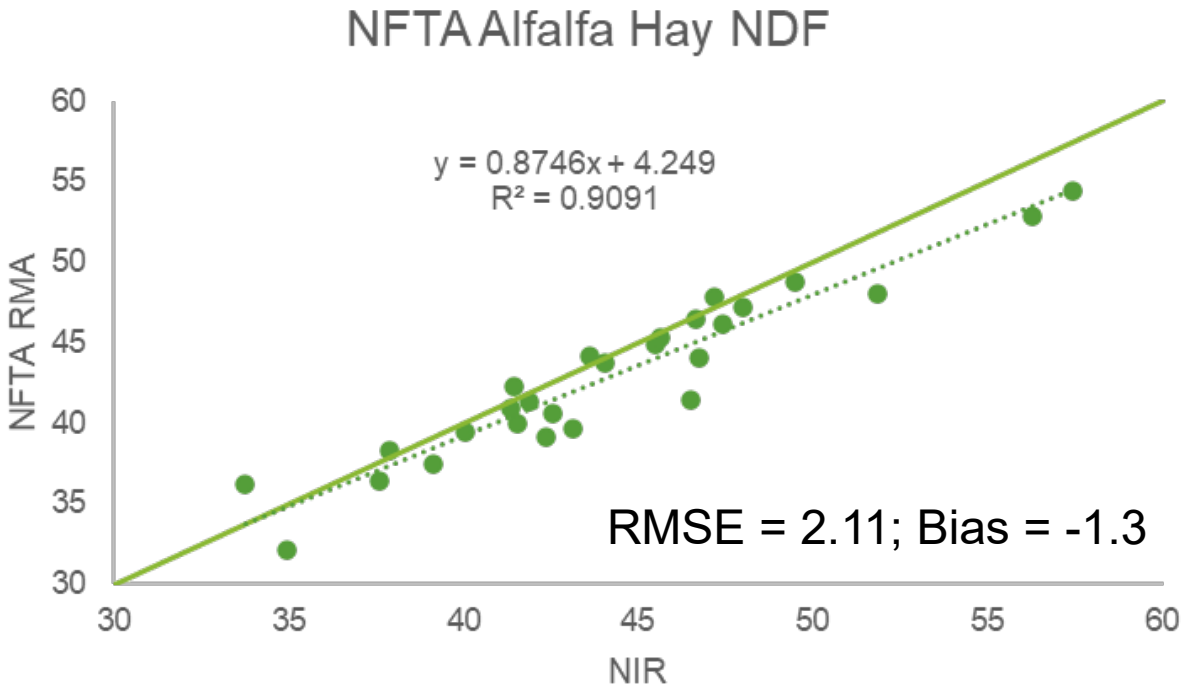
1) Does the calibration work on my instrument?
    1) Yes, we standardize, but total, exact agreement between instruments is impossible.
2) Does the calibration work with my sample prep method?
    1) NIR is a secondary method. Sample preparation (grinder style, grinding size, drying method, drying temperature) affects the NIR spectrum in ways that may influence results.
3) Are my samples inherently different in a way that makes accurate, unbiased predictions not possible?
    1) Are samples from my region different than other regions?
    2) Are my samples “special?”

DAIRYLAND
Laboratories, Inc.

Validating Standardization Efficacy



After standardization, it is common to use samples like NFTA with known reference values as a check on the host instrument to ensure that model performance matches the master instrument.

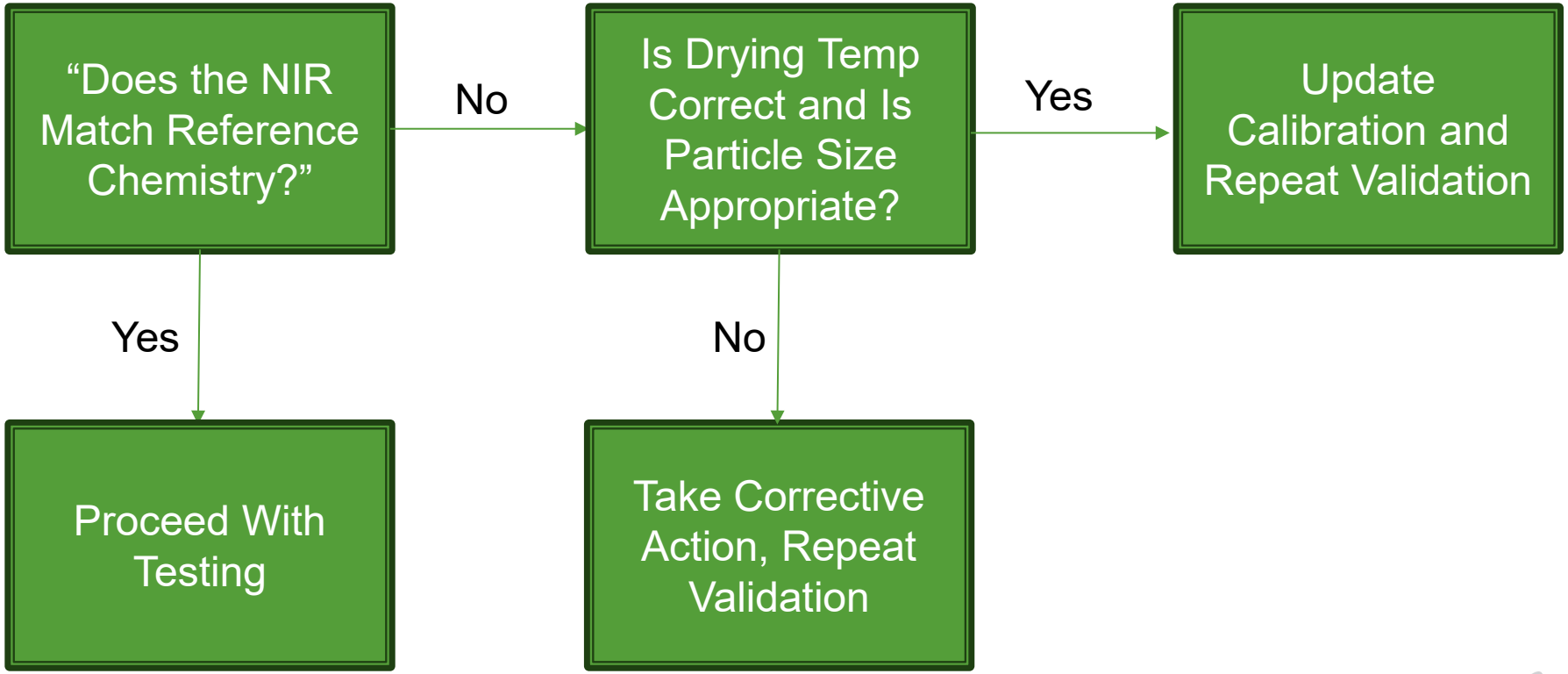


Poor validation results can trigger a redo of the standardization process.

DAIRYLAND
Laboratories, Inc.

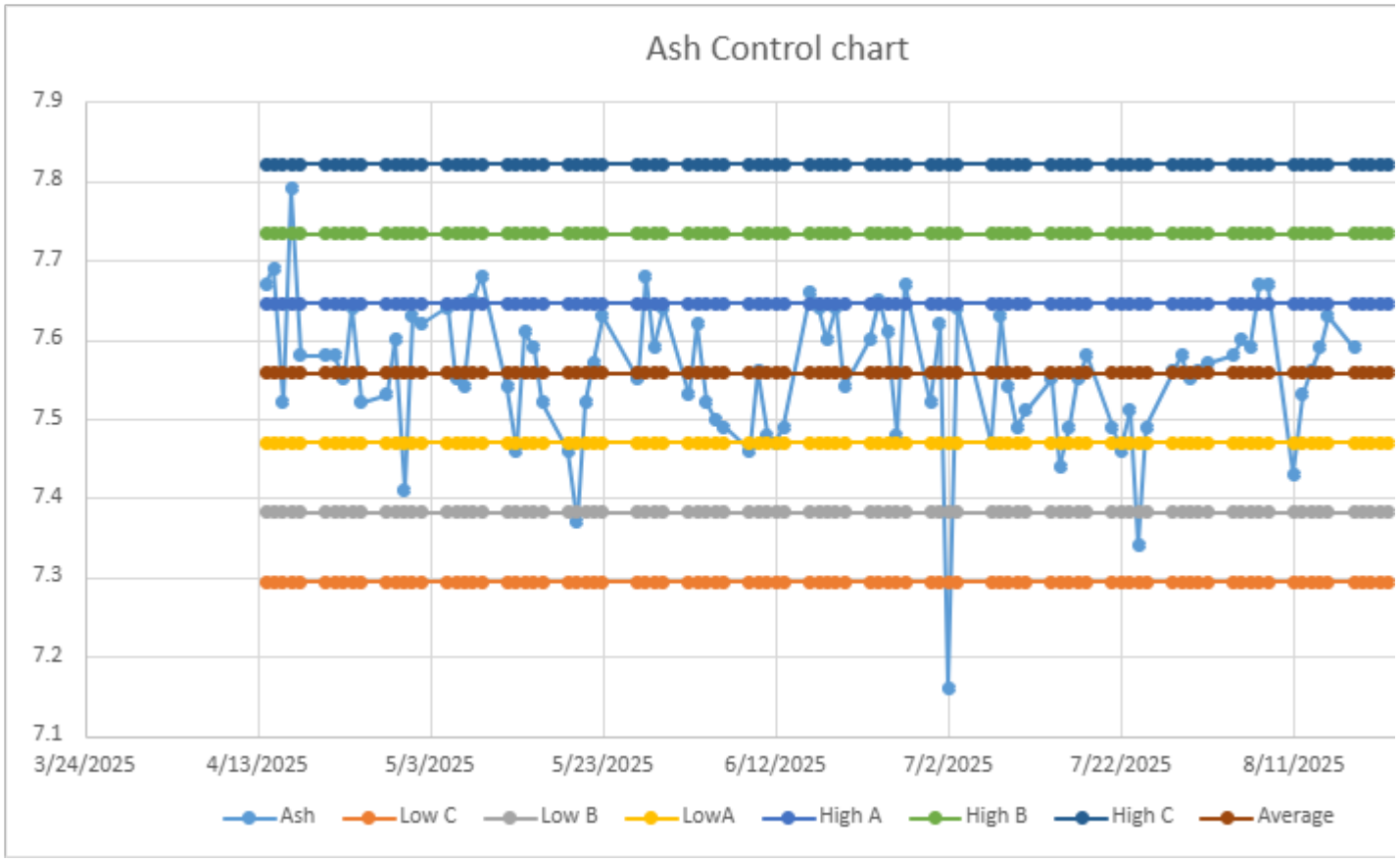
Validating on New User's Samples

After completing standardization validation with known samples, it is good practice to solicit a validation set from a new location and their instrument and following the following procedure:



DAIRYLAND
Laboratories, Inc.

Not All Chemistry is Created Equal…



Mean = 7.56%, Standard Deviation = 0.08%

In some cases, the reference chemistry you are validating against is very precise.

For many simple procedures, the cost of a duplicate rep exceeds the benefit.

DAIRYLAND
Laboratories, Inc.

Repeatability a Concern For Biological Assays

Ross UCP%CP – Blood Meal

	UCP%CP Rep 1	UCP%CP Rep 2	UCP%CP Rep 3	Standard Deviation
1663370	10.00	15.68	9.10	3.57
1663371	**23.42**	**20.60**	**12.14**	5.87
1663373	0.93	1.11	0.53	0.29
1675785	0.55	1.03	0.42	0.32
1686315	**10.75**	**17.95**	**1.75**	8.12
1722283	3.08	5.71	2.95	1.56
1741732	0.88	1.36	0.36	0.50
1741733	11.92	13.03	10.43	1.31
1741734	0.65	4.69	0.00	2.54
1762864	**0.44**	**24.02**	**3.17**	12.90

Average Standard Deviation = 3.70%

Achieving precision on some biological assays or sample types may be more challenging.

In these cases, it may be more wise to select a smaller group of sample to run in duplicate or triplicate for better validation.

DAIRYLAND
Laboratories, Inc.

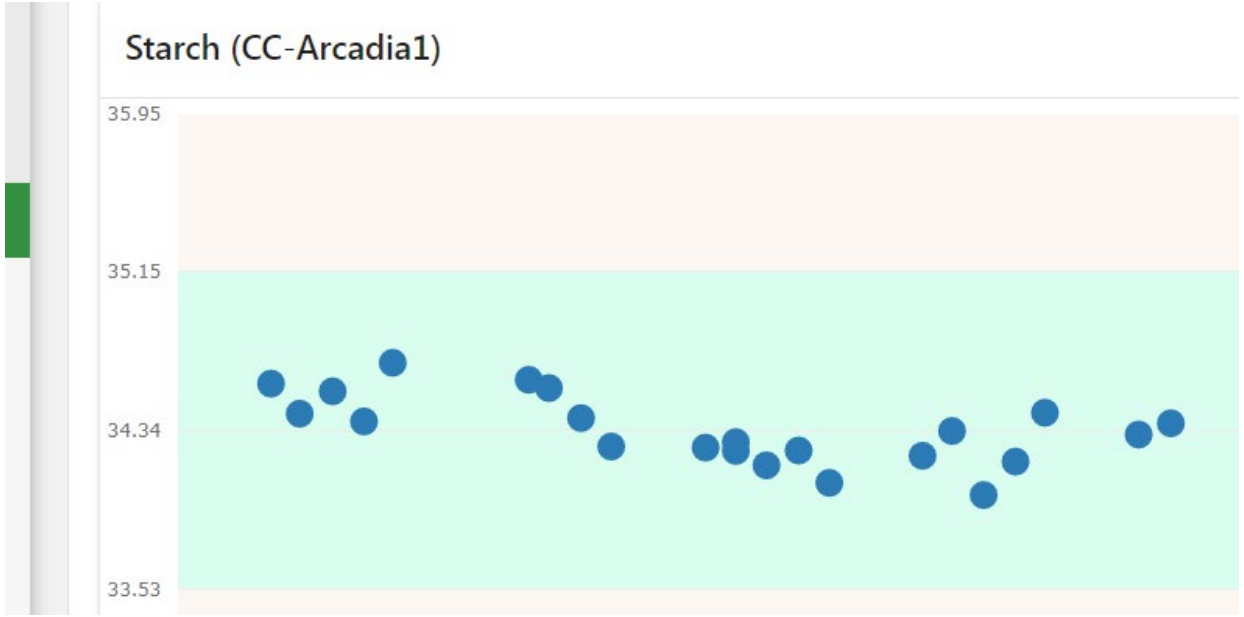
That's How Your Provider Validates…

Is It Good Enough For Me?

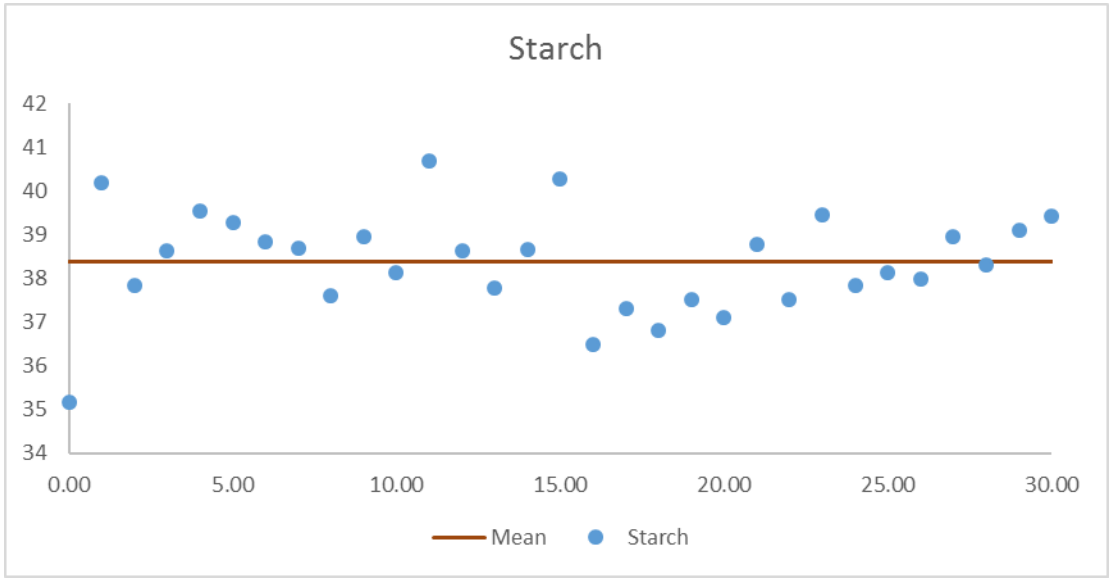
DAIRYLAND
Laboratories, Inc.

NIR vs Reference Chemistry on Control Chart

NIR



Reference Chemistry



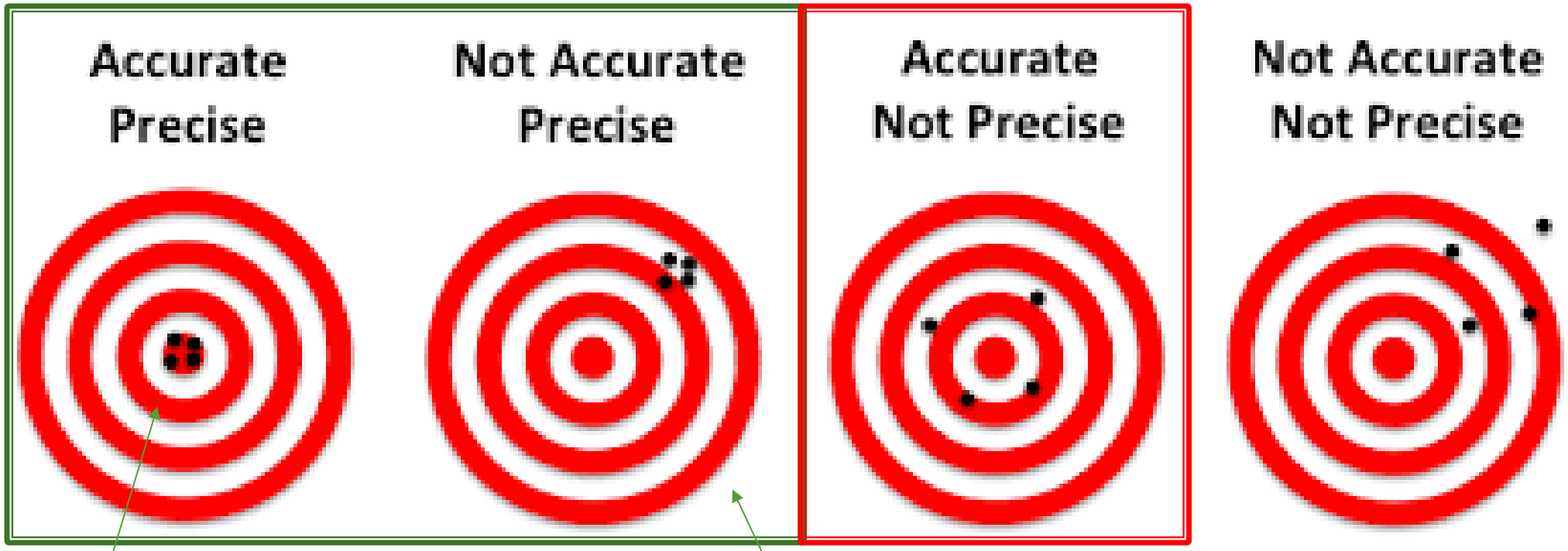
Range = 34.01 - 34.68%, 0.67%

Range = 35.18 – 40.68%, 5.5%

NIR predicted values are very *precise!*

DAIRYLAND
Laboratories, Inc.

Precision Vs. Accuracy



Which of these is NIR?

If carrying out the validation described earlier shows that NIR is accurate (no slope/bias problem) and the control chart shows it is precise, you don't need to do any more validation.

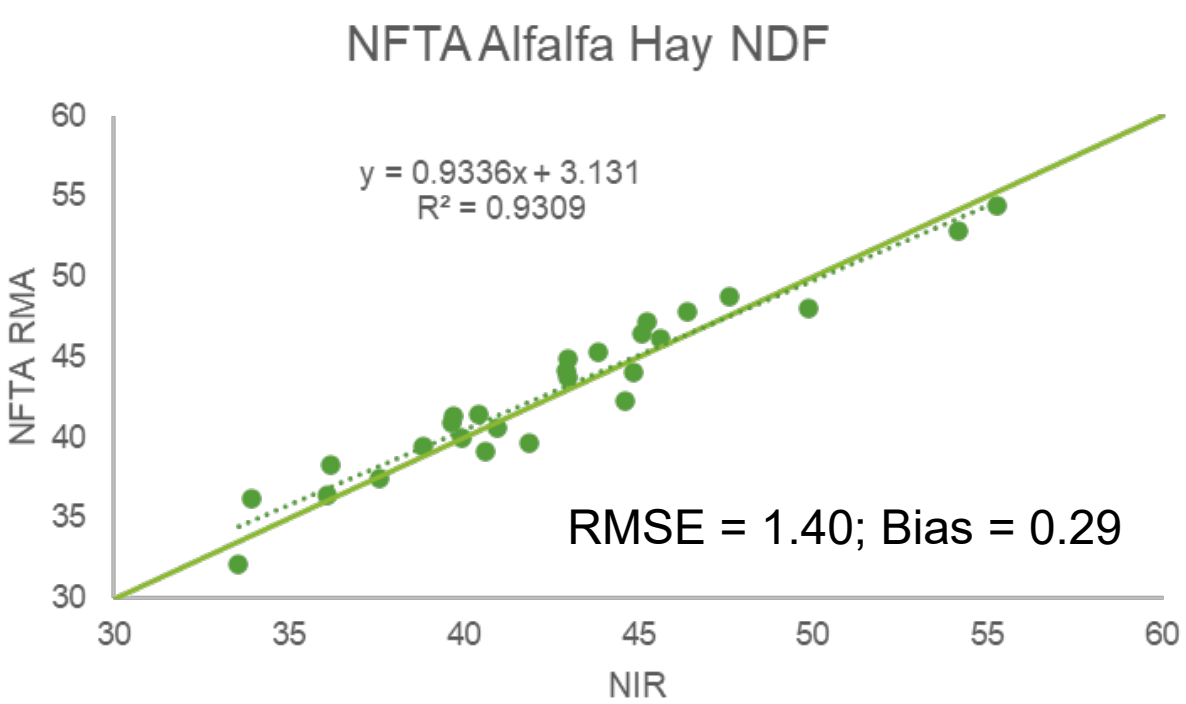
Reference Chemistry

If there is potential for this, a systematic error, you should entertain the thought of validating.

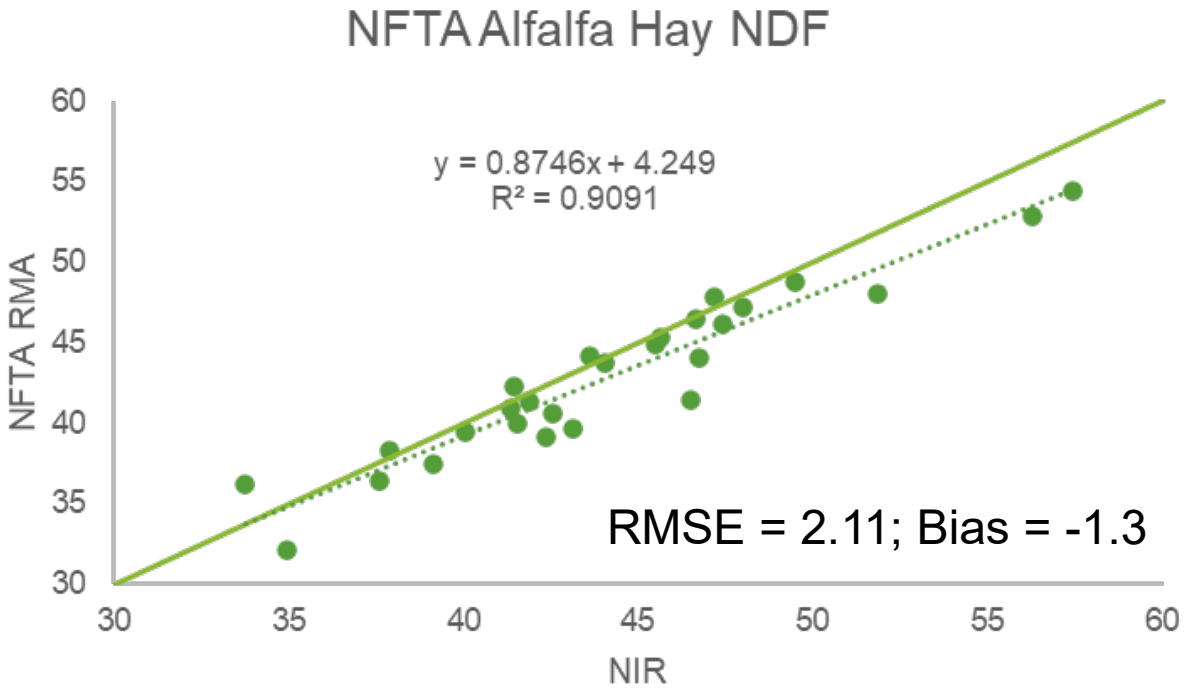
Can we find systematic NIR error?

DAIRYLAND
Laboratories, Inc.

Validating Standardization Efficacy



After standardization, it is common to use samples like NFTA with known reference values as a check on the host instrument to ensure that model is effective on the instrument.

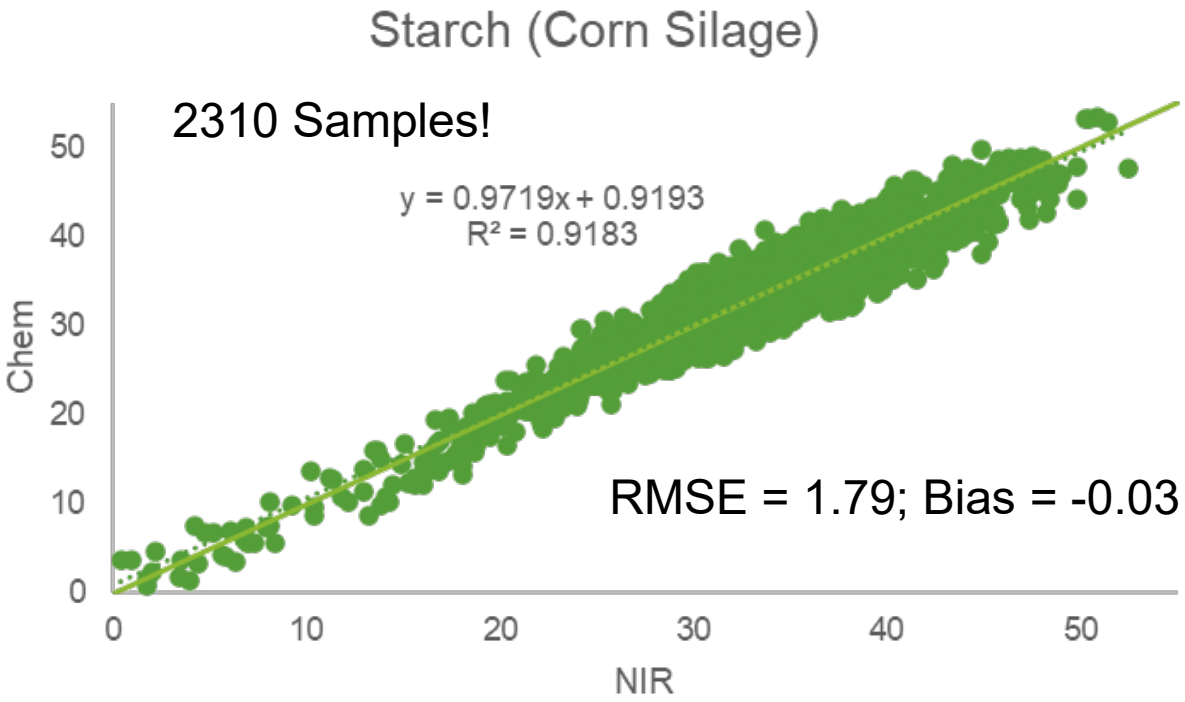


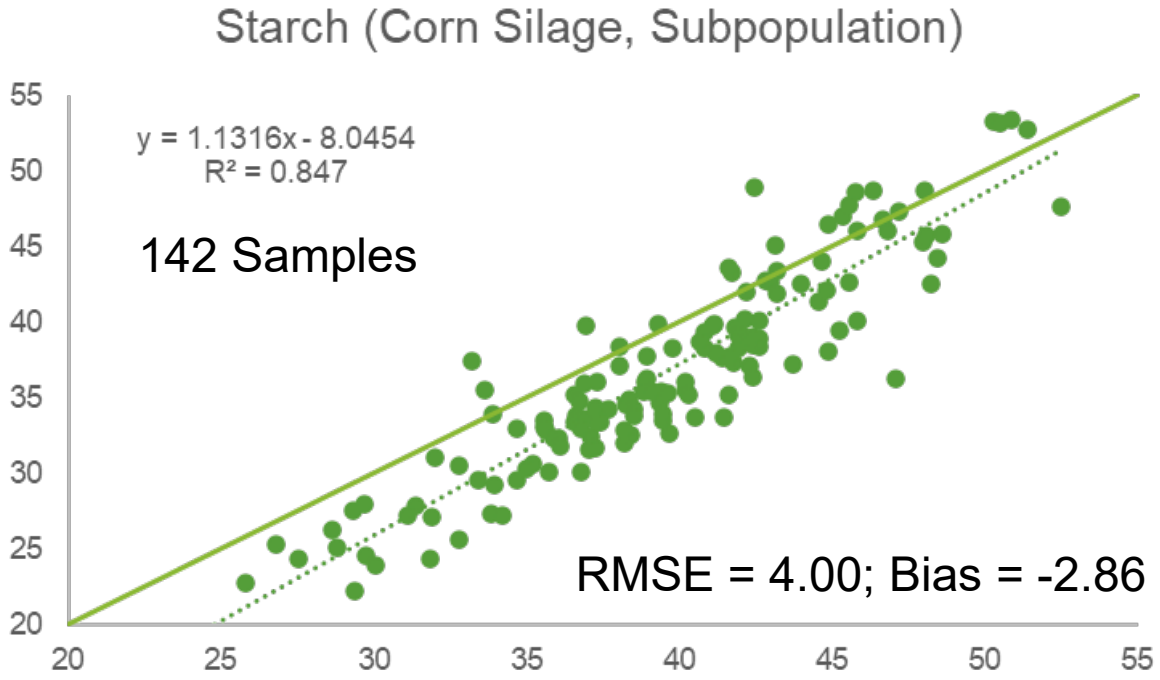
Poor validation results can trigger a redo of the standardization process.

A BIAS ON AN INSTRUMENT IS A SYSTEMATIC ERROR!
(Applies to all Samples)

DAIRYLAND
Laboratories, Inc.

Systematic Error: Example 2





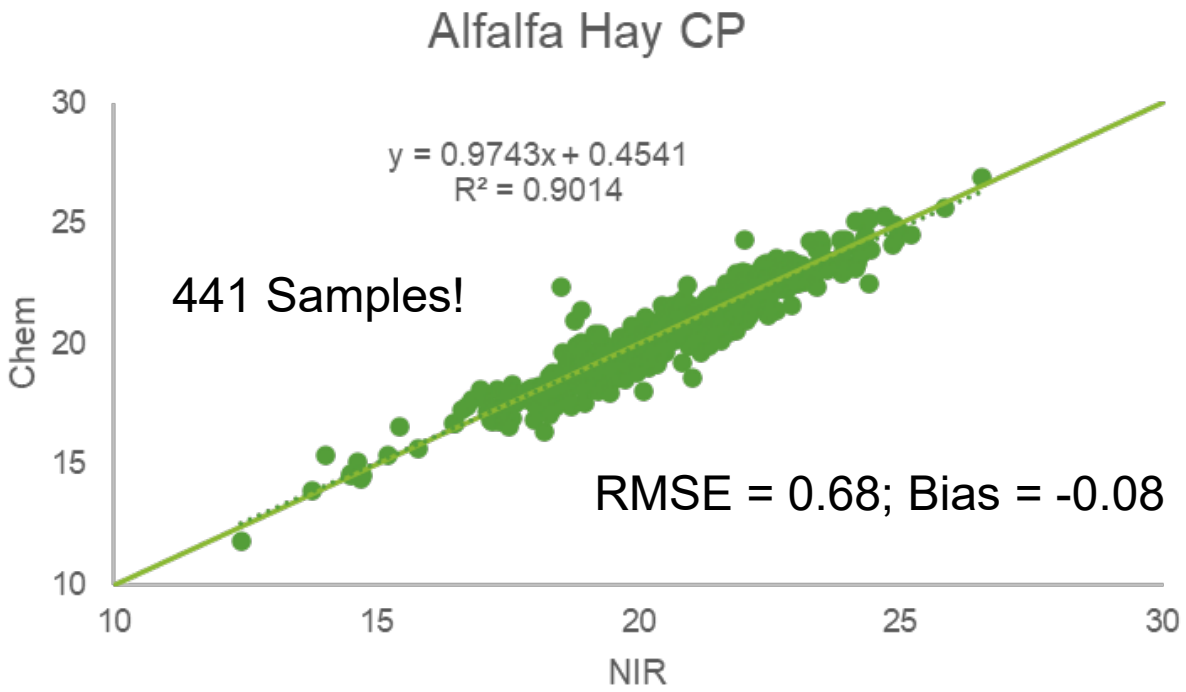
Subpopulation = Samples with RDM <92%

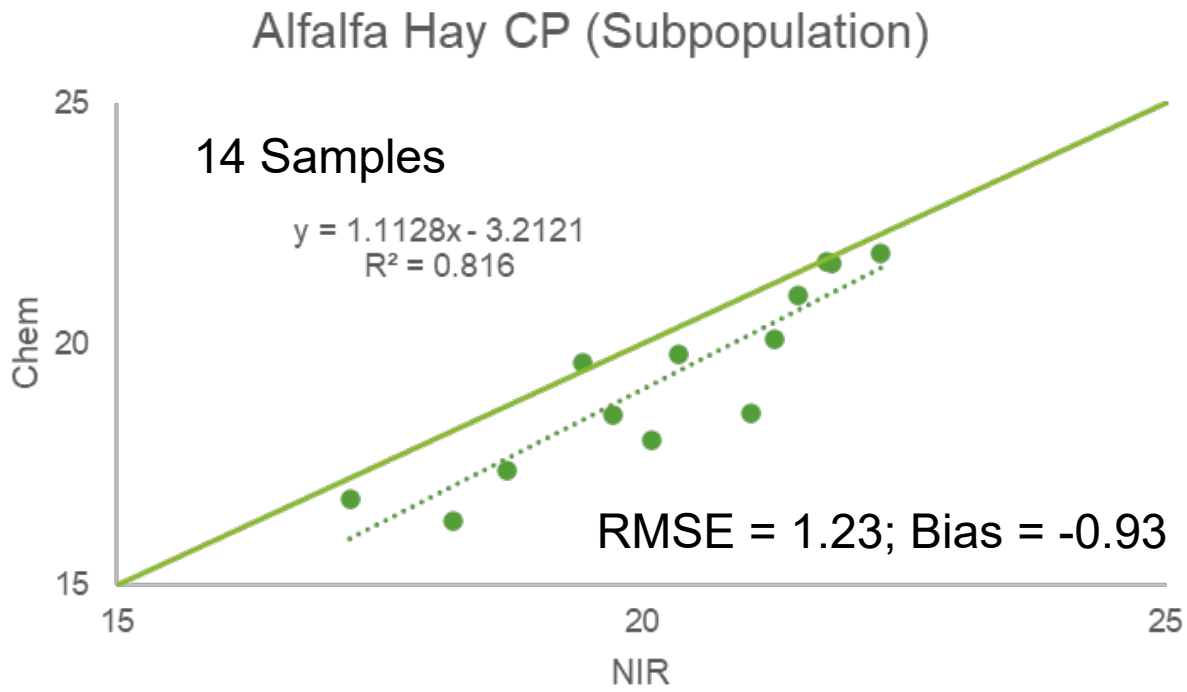
Some would say: “Dry your samples better!” (I agree)

I also say: “There is a systematic over-prediction of starch in corn silage when one chemical component (water) is in abnormal excess.”

DAIRYLAND
Laboratories, Inc.

Systematic Error: Example 3





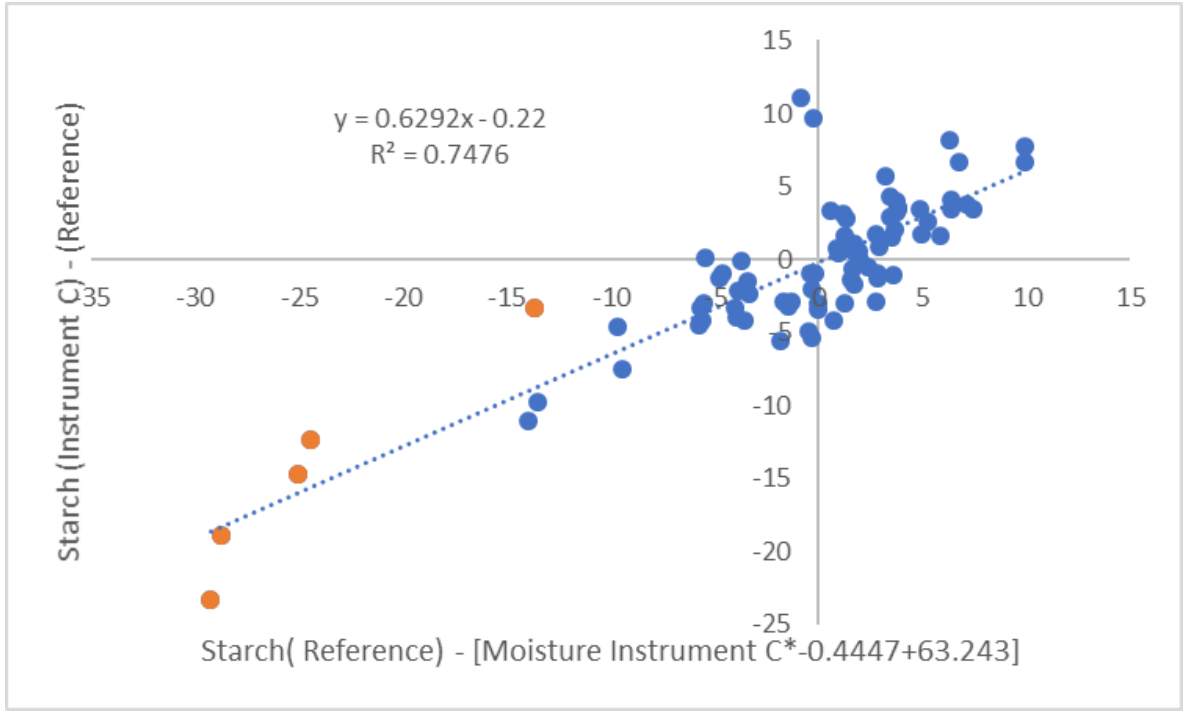
Subpopulation = Samples with Chem Ash > 16%

When ash is high, crude protein may be biased high.

DAIRYLAND
Laboratories, Inc.

Systematic Error in Starch – In-Field NIR

Sample	Reference Moisture	Reference Starch (%)	Predicted Starch (%)	Reference - Predicted Starch (%)
786751_0	79.34	3.52	15.9	-12.38
786750_5	79.92	2.63	17.32	-14.69
786749_0	66.83	4.78	23.63	-18.85
786900_5	70.11	18.38	21.72	-3.34
785907_5	65.92	4.67	27.95	-23.28



DAIRYLAND
Laboratories, Inc.

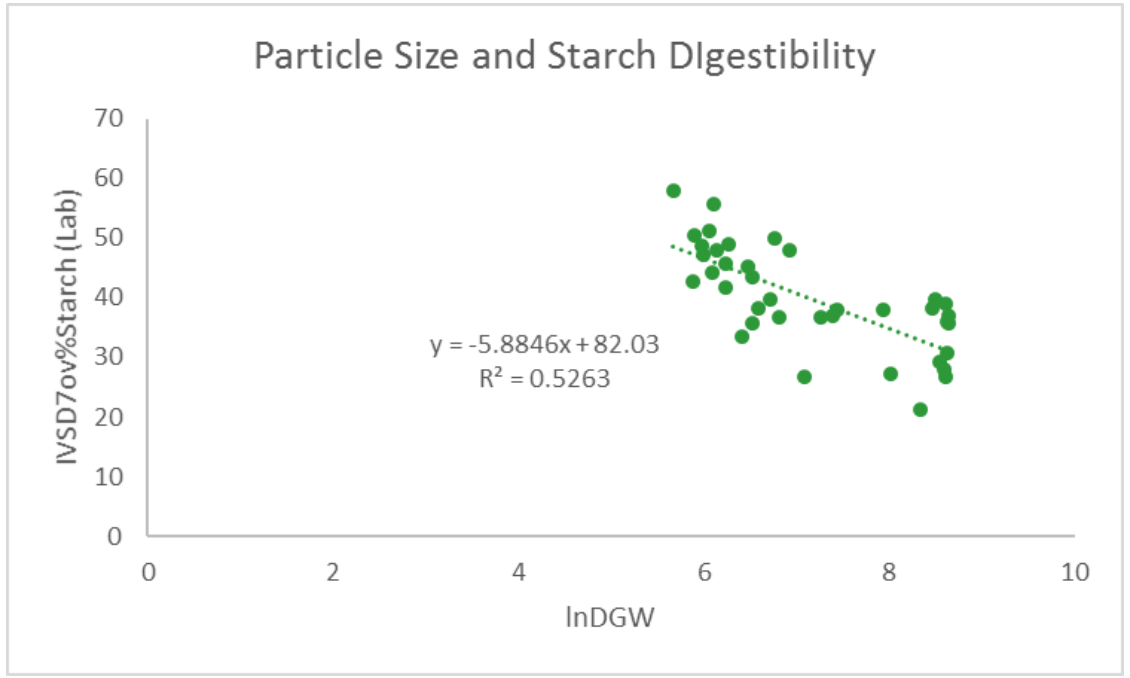
How this is explained:

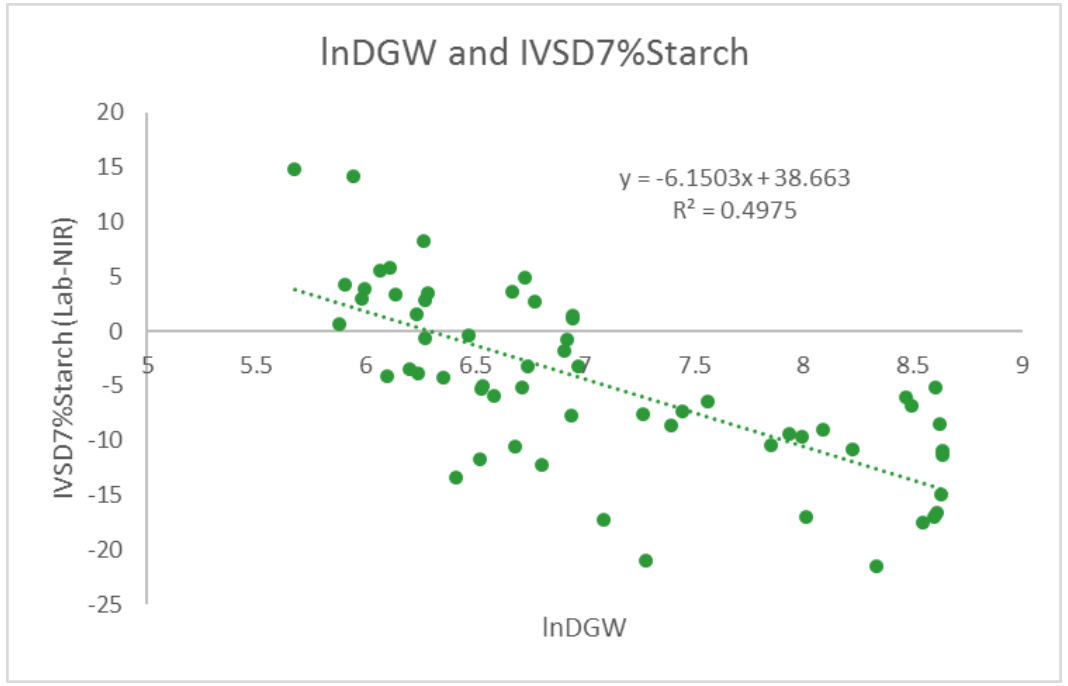
Starch Increases, Moisture Decreases

- As corn matures, moisture decreases and starch increases – they are correlated, but not perfectly.
- In the calibration set: $Starch = -0.4447 \times Moisture + 63.243$

DAIRYLAND
Laboratories, Inc.

In Vitro 7-Hour Starch Digestibility (Dry Corn)





Particle Size and Starch Digestibility are correlated, but we grind the sample and destroy the necessary information needed to solve the problem. Particle size at scanning and original particle size are not correlated.

Error on IVSD7%Starch in corn grain is correlated to the mean particle size of the raw sample.

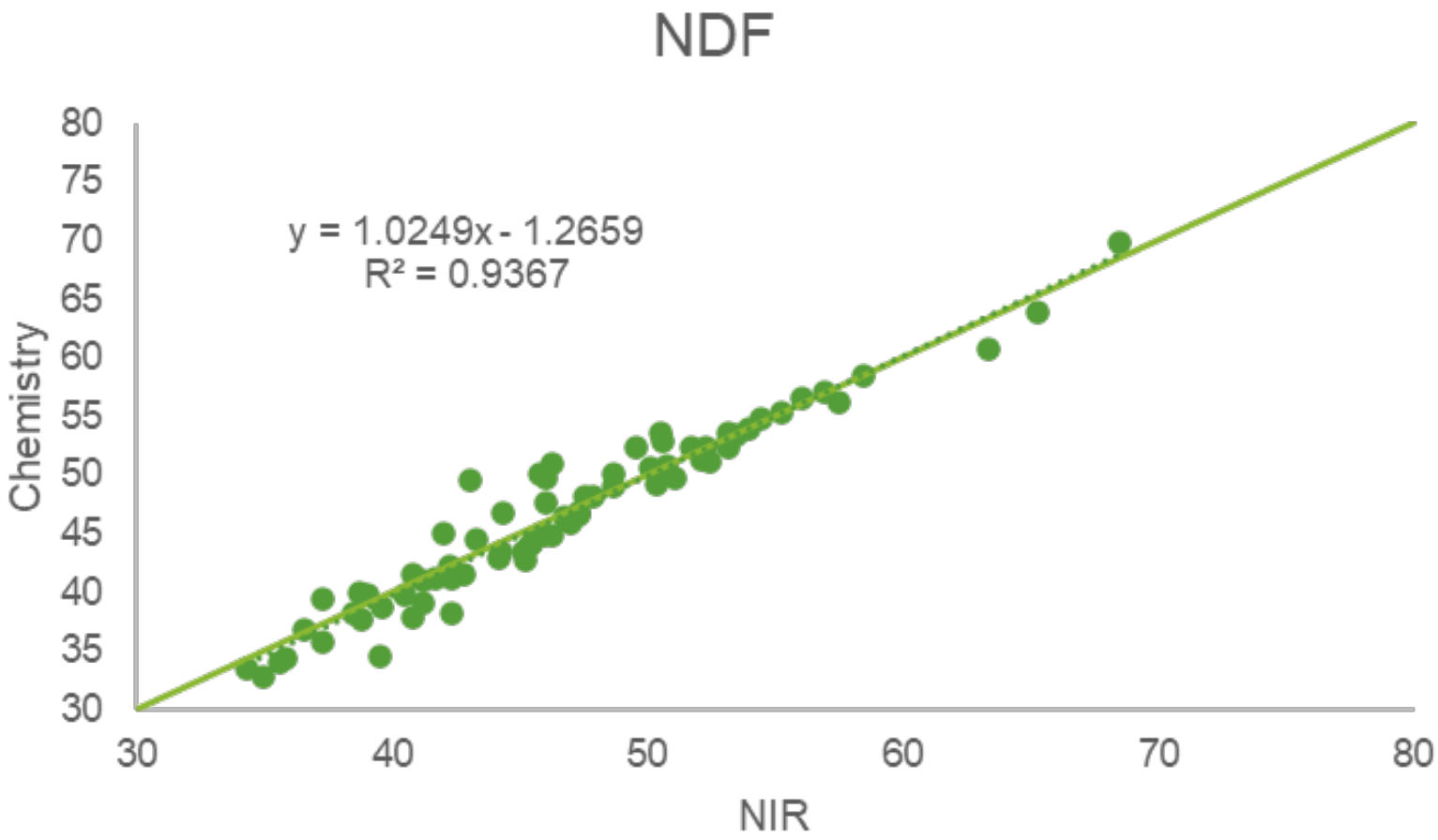
DAIRYLAND
Laboratories, Inc.

Can’t We Just Do things Differently To Remedy the Error?

- For alfalfa hay NFTA set: just redo the standardization.
- For corn silage starch: please dry your samples better.
- For alfalfa hay CP: avoid soil contamination when you sample on the farm.
- In-field NIR starch: we dry and grind samples for a reason.
- Starch Digestibility: perhaps try on unground samples instead?

DAIRYLAND
Laboratories, Inc.

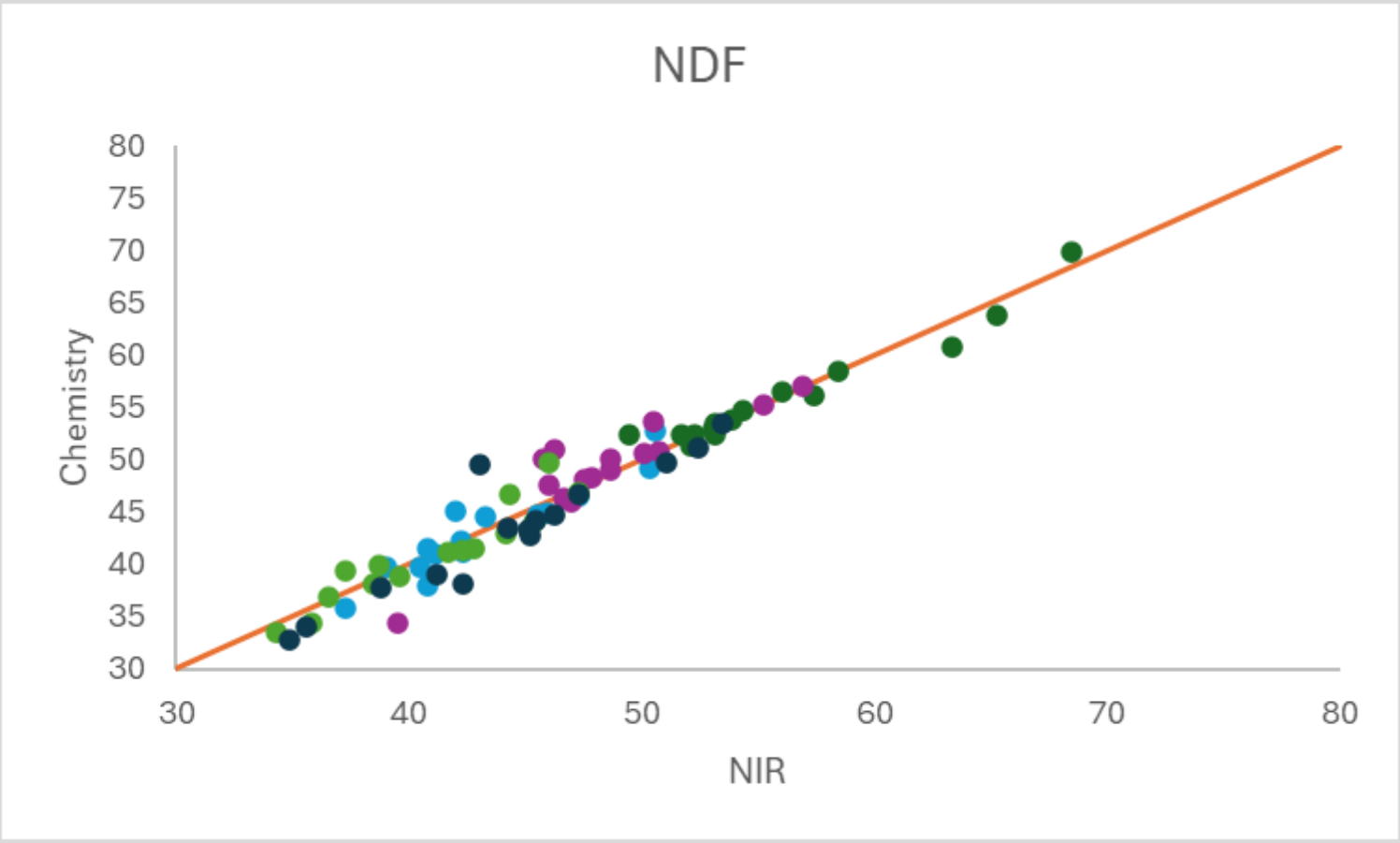
Is This Sorghum Hay NDF Calibration Good Enough?



Validation Set: RMSE = 1.87, Bias = -0.39

DAIRYLAND
Laboratories, Inc.

Sorghum Hay By Variety



For NDF:

Variety	RMSE	Bias	R^2	Slope
1	1.22	-0.07	0.95	0.92
2	1.5	-0.16	0.88	1.05
3	2.33	0.72	0.81	1.13
4	2.44	-1.05	0.86	1.05
5	1.55	0.03	0.88	1.06

For NDFD30:

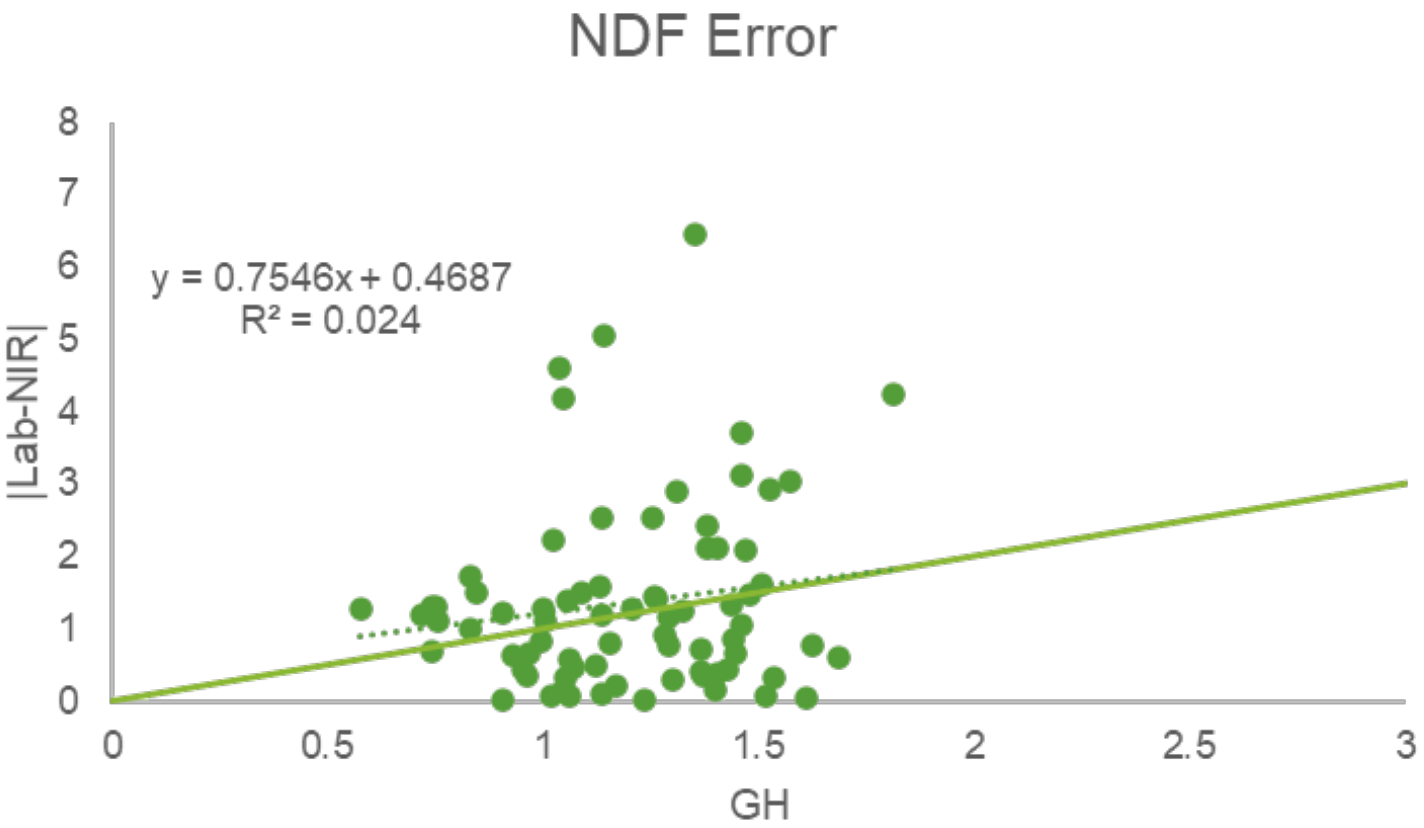
Variety	NDFD30(Lab)	Chem Rank	NDFD30(NIR)	NIR Rank
1	50.49	2	50.84	3
2	49.28	3	46.64	5
3	49.24	4	52.87	1
4	44.83	5	46.85	4
5	55.62	1	51.36	2

Sorghum Hay NDF error is not consistent across variety!

This can propagate into larger error for calculated values!

DAIRYLAND
Laboratories, Inc.

Can H Values Help ID This?



GH value is NOT correlated with magnitude of error!

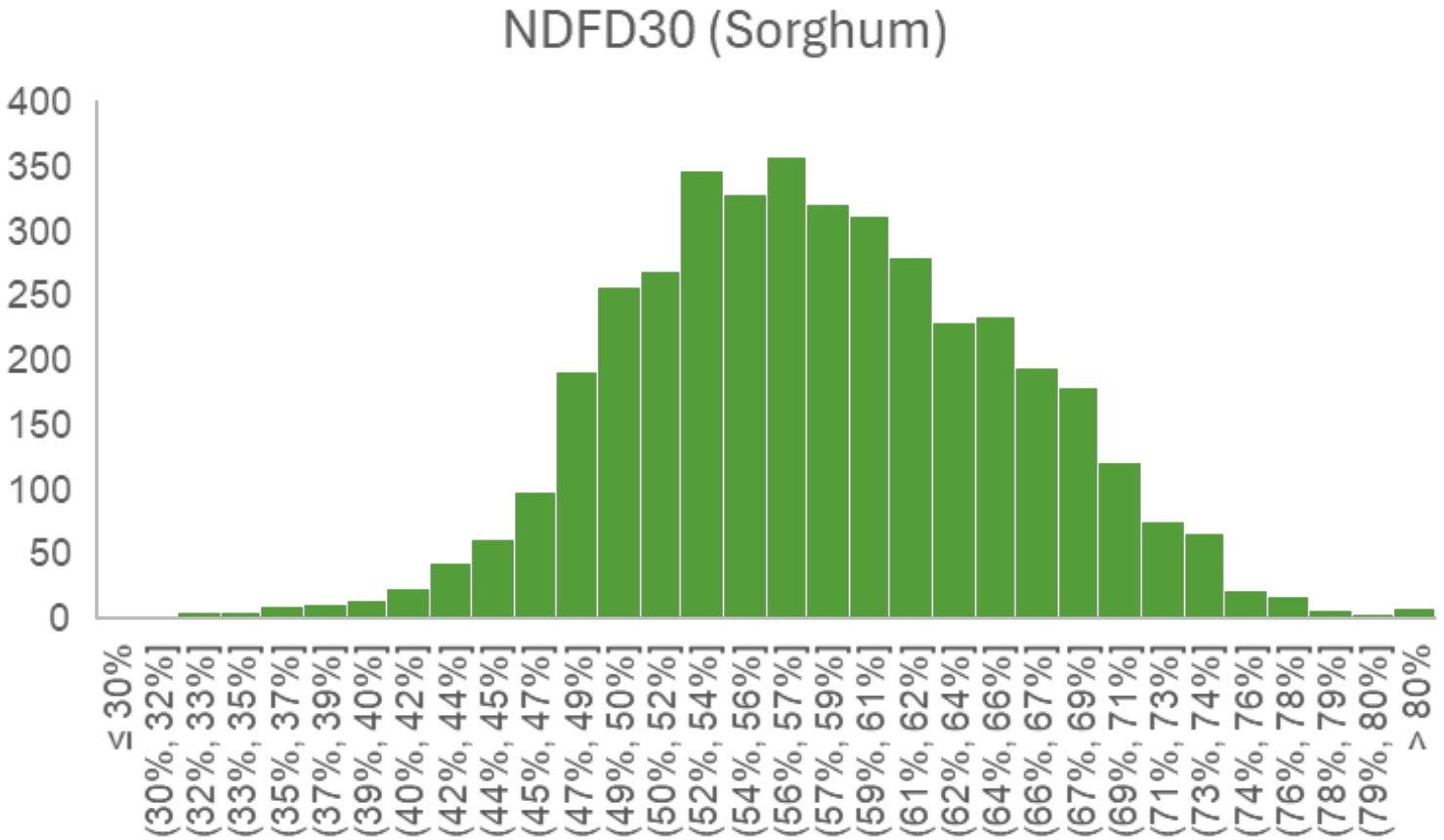
GH by Variety:

Variety	Mean GH	RMSE	Bias
1	1.2	1.22	-0.07
2	1.28	1.5	-0.16
3	1.29	2.33	0.72
4	1.08	2.44	-1.05
5	1.22	1.55	0.03

GH > 3 is considered an outlier – these aren't close!

DAIRYLAND
Laboratories, Inc.

In the Context of the Population…



Mean = 57%, Standard Deviation = 8%

The biases in NDFD by variety is small compared to the variation of the population...

For NDFD30:

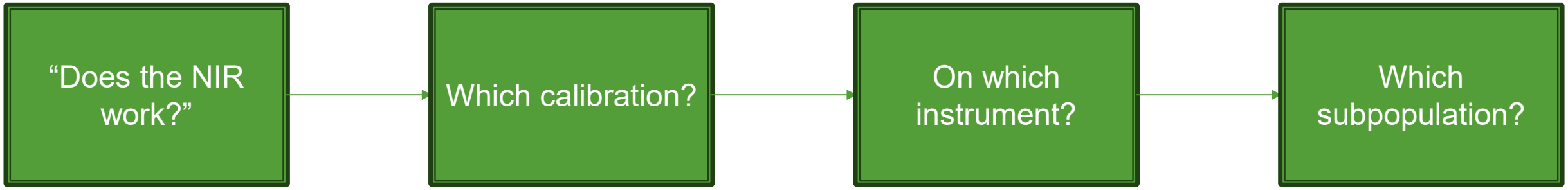
Variety	NDFD30(Lab)	Chem Rank	NDFD30(NIR)	NIR Rank
1	50.49	2	50.84	3
2	49.28	3	46.64	5
3	49.24	4	52.87	1
4	44.83	5	46.85	4
5	55.62	1	51.36	2

… but the biases in NDFD are significant when making comparisons between varieties.

DAIRYLAND
Laboratories, Inc.

Tie-Up From Last Example

- If we repeated this exercise with new samples from same varieties, would we reach the same conclusion? I don't know.
- Someone did this study, and probably drew a conclusion from it.
- "The calibration provider will validate on EVERYTHING, not just your samples!"



There are ~640 nutrient/feedtype combinations at Dairyland that have a unique calibration.

We have 139 standardized NIR instruments in our system.

Region, high/low, fresh, variety/crop type, nutrient relationship, etc. ("10")…

640 X 139 x 10

889,600

Sample Info We Would Need

Sample Info We Get

In Summary…

- You cannot judge performance of NIR models from looking at calibration statistics – you must validate (compare reference chemistry and NIR predicted values)
- A reputable provider of NIR calibrations will have validated, and there will be no systematic errors (slop/bias issues) for the calibrations *across the entire tested population,* and probably a few obvious subpopulations.
- Depending on what your goal is and what resources you have, you may want to validate for your specific use case to verify that no systematic error exists within your subpopulation which may affect your findings or decision.

DAIRYLAND
Laboratories, Inc.

Thank you for your attention…

Questions?

DAIRYLAND
Laboratories, Inc.