Farm to Fridge: Digital Traceability and Quality Upgrading in the Kenyan Dairy Value Chain*

Guanghong Xu†

Job Market Paper

November 17, 2024

Please click here for the latest version

Abstract

Disorganized agricultural value chains often prevent the transmission of quality incentives to upstream farmers, especially when quality is costly to observe at the farm gate. This study develops digital traceability systems for Kenyan dairy cooperatives and introduces an innovative quality monitoring method, using Bayesian statistical models to infer individual milk quality from pooled samples and reduce testing costs. I reveal randomly selected farmers' milk quality either by the model or by random tests to both cooperatives and farmers. I find that model group farmers reduce added water by a significant 21.9% compared to the control group, while random test group farmers show an insignificant 12.6% reduction. Additionally, treated farmers who consistently deliver high-quality milk get higher credit limits from the cooperative and use more credit for animal feed. A back-of-the-envelope calculation suggests that the model offers cost-effective quality improvement, making it a promising tool for continuous quality monitoring and scalable implementation.

JEL Codes: C11, D82, O12, O13, Q12, Q13

Keywords: Traceability, Quality Upgrading, Moral Hazard, Bayesian Estimation

*I am heavily indebted to Jonathan Robinson, my advisor, for his mentoring and support throughout the life of this project. I am also sincerely grateful to Alan Spearot, Ajay Shenoy, Jessie Li, Shilpa Aggarwal, Susan Godlonton, Natalia Lazzati, Gerelt Tserenjigmid, David Schönholzer, Ariel Zucker, Nils Teufel, and James Rao for their tireless support. I also thank seminar participants at UCSC and ILRI for their helpful discussions. I am grateful to Steve Biko for outstanding fieldwork assistance and to ILRI Kenya for administrative support. Funding for this project was generously provided by the National Science Foundation (NSF), MIT/J-PAL Digital Agricultural Innovations and Services Initiative (DAISI), Weiss Family Fund for Research in Development, CGIAR Standing Panel on Impact Assessment (SPIA), SurveyCTO, UCSC Blum Scholars Grant, and UCSC Economics Dissertation Research Grant. The experiment and data collection was approved by the UCSC IRB, the International Livestock Research Institute (ILRI) Institutional Research Ethics Committee (IREC), as well as the in-country research permit from the National Commission for Science, Technology, and Innovation (NACOSTI). This trial is registered on AEA's registry for RCTs (AEARCTR-0014255). All errors are my own.

†Department of Economics, University of California, Santa Cruz. Email: guanghongxu@ucsc.edu. Website: https://guanghongxu.github.io/.

1 Introduction

Improving the quality of agricultural products is often considered an important step in the process of structural transformation, moving farmers from subsistence farming to market-driven, commercial agriculture. However, disorganized supply chains with many layers of intermediaries often prevent the transmission of quality incentives to upstream farmers, especially when quality is unobserved at the farm gate, which discourages farmers from upgrading quality. A growing literature finds that quality certification can reduce this market friction caused by asymmetric information and improve product quality (Bernard et al. 2017; Macchiavello and Miquel-Florensa 2019; Abate et al. 2021). The cost to get certified, however, can be high and not financially feasible for smallholder farmers (Bergquist and Startz 2024), or even farmers' cooperatives (Aggarwal et al. 2024b). Therefore, in markets where the producers are mainly smallholders and products are aggregated by intermediaries, downstream buyers typically certify the quality in bulk after aggregation. Although studies show that producers respond to quality incentives and improve product quality (Atkin et al. 2017; Hansman et al. 2020; Magnan et al. 2021; Bold et al. 2022; Hoffmann et al. 2023; Park et al. 2023; Rao and Shenoy 2023; Deutschmann et al. 2024), when quality is only certified further downstream in the value chain, whether and how the quality incentives are transmitted to upstream producers remain open questions.

This paper explores a new potential solution to enhance accountability and incentivize quality improvement in value chains: establishing a traceability system that enables precise rewards for farmers who deliver high-quality products, particularly when quality is difficult and costly to observe at the farm gate. Focusing on the Kenyan dairy value chain, this study examines the impact of introducing digital traceability systems among dairy cooperatives. I test a novel quality measurement system derived from the digital traceability system, comparing it to random farm-gate testing and a control group. I find that cooperatives and farmers respond to both quality monitoring methods, and the milk quality improves at the endline. The digital traceability system outperforms conventional random testing in most metrics. Additionally, farmers in both treatment groups who consistently provide high-quality milk get a higher credit limit from the cooperative and use more credit for animal feed at the endline.

In the Kenyan dairy (formal) value chain, farmers are selling their milk to dairy cooperatives. Dairy cooperatives hire milk transporters to collect milk from farmers. Milk is not tested at this point due to the prohibitively high testing cost (both time and monetary), and thus, farmers are paid based solely on quantity. Transporters usually aggregate the milk from multiple farmers to fill larger milk cans, which are then transported to collection centers. The collection centers pour the milk together into cooling plants (usually containing 10 to 100 of these cans). Cooperatives then sell the aggregated milk to processors who either

accept it with a premium price, accept it without a premium, or reject it based on their comprehensive milk testing results.[1] If the milk is rejected, the entire vat of milk has to be destroyed by law, and the dairy cooperative will bear the loss.[2] Therefore, quality incentives exist in the downstream markets but are not transmitted to the upstream farmers, which leads to a low-quality, low-price equilibrium.

To effectively address this market failure and evaluate testing schemes that mitigate it, it is important to recognize that milk quality defects have both farmer-specific and time-varying components. To be specific, milk processors check water adulteration, butterfat content, and bacterial load when buying milk in the study region. Water adulteration is purely a day-to-day choice, and different farmers have different tendencies to add water to the milk. Butterfat is determined by cow breed type (fixed) and animal feeds (varying), and it could also be affected by added water (varying). Lastly, for bacteria, some farmers might not have the knowledge (fixed) of milking hygiene, while others, even if knowledgeable, find it costly (varying) to follow the recommended milking hygiene protocol. If milk quality is solely determined by fixed, farmer-specific characteristics, a one-time test could effectively reveal this information. However, in this setting, milk quality also depends on variations in farmer behavior, indicating that a one-time test would be insufficient to fully capture these dynamics.

In the context of dynamic moral hazard problems with limited monitoring resources, random monitoring is largely studied in the existing literature (Jin and Leslie 2003; Varas et al. 2020; Ball and Knoepfle 2024). Alternatively, this paper examines a new approach to still monitor daily, but at the aggregate level.[3] In principle, by testing aggregated milk cans with recorded information on which farmers contributed to each can, I could infer individual milk quality. Since the composition of each container varies daily, the can-level testing outcomes would reflect these differences, enabling individual-level assessments.

To monitor farmers' milk quality every day, I developed a digital traceability system to track the milk throughout the value chain so the system can provide a basis for continuous monitoring. This traceability system helps to record who contributes to each aggregated milk can and their corresponding contribution percentages. In addition, I test these aggregated cans when they arrive at the cooling plant for several days. Using the aggregated can-level testing data and can composition data, I developed novel Bayesian statistical models to infer

[1] The reason is that high-quality milk can be used to produce high-value products like buttermilk and yogurt. Medium-quality milk is limited to use in ultra-pasteurized liquid packets, but low-quality milk cannot be traded by law.

[2] Note that this extreme case exists, but is uncommon.

[3] Pooled group monitoring is also commonly used during the severe COVID time back in 2021 and 2022, especially in China, when testing is needed every day until all the positive cases are identified in the whole community.

whether individual farmers produce high- or low-quality milk. The model-predicted quality shows a high correlation with the one-time milk test among 940 dairy farmers in rural Kenya.

To test both the random monitoring and model detection in this dynamic moral hazard setting, I run an individual-level randomized controlled trial with 940 farmers from cooperatives in two counties in Kenya to provide randomly selected farmers' milk quality information either by the model or by individual tests to dairy cooperatives, as well as to farmers themselves. I find that cooperatives react to the quality information provided to them; they train low-quality (butterfat and bacteria) farmers and talk to farmers about water adulteration. Both cooperatives chose to have a conversation with low-quality farmers first without imposing any real consequences during the first two rounds of quality monitoring and information sharing. After the third round, a small number of farmers who ignored the warning and continued the practice of diluting milk with water had a portion of their payment temporarily withheld.[4] As a result, model group farmers had a significant 21.9 percent reduction in added water compared to the control group, and individual test group farmers had an insignificant 12.6 percent reduction. A back-of-the-envelope calculation indicates that this model delivers a greater reduction in added water per dollar spent on testing. Moreover, it reduces the time required to collect individual milk samples in the field and provides can-level quality information to distinguish good milk from bad milk before mixing occurs at the cooling plant. These benefits highlight the potential of this new quality monitoring approach as a superior method for continuous milk quality monitoring and regulation.

I find no significant effect on bacteria and the absolute level of butterfat. One possible reason is that the baseline level of butterfat is already high, and even though I do not find significant effects on the average butterfat level, there are significantly 11% and 10% fewer farmers failed the minimal required national standards of butterfat in the random tests group, and model group correspondingly, compared to the control group. Training alone without incentives seems only to reduce the probability of farmers failing the minimal required national standards of butterfat and has no effect on the absolute butterfat level and bacteria. This is consistent with Park et al. (2023) who find that training helps in quality upgrading for dragon fruit farmers in Vietnam, and its effects are much larger when accompanied by incentives from exporters. In fact, I observed a significant reduction in the probability of farmers failing the minimal required national standards of bacteria after the second round of information sharing, but the effect disappeared after the third round. This suggests that cooperatives will need to provide clear quality incentives to farmers to have a long-lasting effect.

In addition, treatment farmers who never add water get a higher credit limit from the

[4] All payments were restored later on; see the detailed discussion in Appendix Section B.

cooperative and use more credit for animal feed, as the cooperative believes that these farmers are more trustworthy and will be more likely to pay back the loans. It appears that the flow of quality information will also facilitate the flow of credit provision along the value chain. Improved access to credit for farmers producing high-quality products creates a positive cycle along the value chain: fostering better quality, increased credit opportunities, and ultimately higher income (Casaburi and Willis 2024).

This paper contributes to several strands of the literature. First, it speaks directly to the growing literature on quality certification. As Akerlof (1970) first pointed out theoretically, asymmetric information about product quality can lead to a "lemons" market, negatively affecting the welfare of market participants. Later on, Viscusi (1978) shows that a system of quality certification can help to mitigate this issue. This study directly engages with the expanding literature on quality certification, including Bernard et al. (2017), which studies the effect of informing villagers about upcoming onion quality certification in Senegal and finds improvements in quality. Macchiavello and Miquel-Florensa (2019) exploits the staggered roll-out of the Quality Sustainability Program (which includes price premium on quality, extension services, and access to inputs for plot renewal altogether) in the Colombian coffee sector to show that eligible farmers produced better quality coffee on average in response to this program. Bai (2024) randomizes stickers to differentiate the quality of watermelons at the sellers' discretion and finds that retailers were able to sell the laser-cut labeled (for higher quality) watermelons to consumers at a higher price. Other ongoing studies include Abate et al. (2022) about wheat certification in Ethiopia; Bergquist and Startz (2024) about quality testing of honey in Ethiopia; as well as Aggarwal et al. (2024b) about cassava cuttings certification in Rwanda. The current study also contributes across this spectrum but is different from the perspective that the quality certification at an aggregate level is already there downstream in the value chain in my study, and quality certification at a disaggregate level would be more like a quality "audit", as it won't certify individual producers' product quality until the aggregate quality is found low.

This study also relates to another strand of literature on buyer-driven upgrading, which links sellers to the market paying a quality premium. This includes Bold et al. (2022), which provides smallholder farmers with training and access to a high-quality maize market in Uganda at the village level, which led to large improvements in quality, productivity, farm gate prices, and thus profits. Atkin et al. (2017) connects rugs producers in Egypt to foreign buyers paying a premium for higher quality rugs at the producer level and finds that producers improved both quality and productivity. Hansman et al. (2020) finds that fish meal exporters in Peru upgraded the quality when access to high-quality export markets was improved. Another ongoing project in Rwanda is also in this area: Aggarwal et al. (2024a) is evaluating the effects of providing farmers with access to maize processors (who

require higher quality standards and pay higher prices than those offered by local markets) on their input decisions. Buyer-driven upgrading can also be achieved by directly incentivizing quality based on existing trading relationships. This includes Hoffmann et al. (2023), which offers an experimental subsidy for farmer groups in Kenya for maize that meets aflatoxin standards and finds a doubling of investment in a bio-control technology. In the dairy sector, Rao and Shenoy (2023) provides collective price incentives for aggregated milk quality at the village level in India and finds an improvement in quality, given that the management team can control the information disclosure. My study is different in the sense that Kenyan dairy cooperatives have already linked to processors paying a premium for higher quality, and I am not increasing the quality premium level in the existing training relationship. Instead, I am focusing on the further transmission of quality incentives to upstream farmers.

To study the transmission of quality incentives to farmers, an ongoing project by Bai et al. (2024) gives quality incentives to coffee traders in Uganda at the parish market level to see whether it can transmit to farmers through market competition in sourcing high-quality coffee cherries. This approach could be effective when quality attributes are at least partially observable (Do Nascimento Miguel 2024). Alternatively, my study focuses on establishing a traceability system along the value chain and evaluates an innovative quality monitoring method based on Bayesian statistical models, which can work when quality is completely unobservable. Bai (2024) finds that one year after the intervention, all retailers stopped sorting and labeling because the cost of the laser technology for an individual seller was prohibitive. Indeed, as Abate et al. (2021) points out as one of the four necessary conditions for quality certification to succeed, the cost-benefit of quality certification is critical to sustainability. A back-of-the-envelop calculation suggests that the model provides a cost-effective way to improve quality, making it a valuable tool for continuous quality monitoring and scalable implementation.

Taken together, the intellectual merit of the study is three-fold. First, different from most quality certification studies, this study would not certify individual producers' product quality until the aggregate quality is found low, which is potentially more suitable in the market where the producers are mainly smallholders, and products will be aggregated by intermediaries. Second, this paper is unique in its focus on the transmission of quality incentives to upstream producers through an innovative quality monitoring method based on traceability systems and Bayesian statistical models, which can potentially reward farmers more precisely for providing high-quality products. Third, this study establishes a quality monitoring system together with the central actor in the market, in my case, the dairy cooperatives. The system is cost-efficient and is likely to be durable and scalable due to the existing long-term and stable relationship between dairy farmers and cooperatives.

The rest of this paper is organized as follows. Section 2 explains the context and traceability

system. Section 3 describes the Bayesian statistical models of detecting individual farmers' milk quality; Section 4 describes the experimental design; Section 5 presents the experimental results; and Section 6 concludes.

2 Context

2.1 Kenyan Dairy Sector

The empirical setting is the Kenya dairy value chain. The dairy sector is one of the largest agricultural sub-sectors in Kenya (KIPPRA 2018); it accounts for more than 4 percent of GDP and 12 percent of the agricultural GDP (KNBS 2019). According to the national survey *Kenya Integrated Household Budget Survey (KIHBS) 2015/16*, among 13,086 households surveyed, 57 percent of them own at least one cow in Kenya.[5]

The national long-term strategy, *Kenya Vision 2030*, has acknowledged that the dairy sector is a key agricultural sub-sector and wants to realize a significant increase in exports of milk and dairy products. Currently, only a small fraction of Kenya's milk production is exported, and a number of trade conflicts have arisen when regional importing countries rejected milk products processed in Kenya in recent years on the grounds that Kenya's raw milk was of insufficient quality. Kenya's raw milk is documented to be of low quality and does not meet national and international standards due to water adulteration, low butterfat, and high bacterial load (Nyokabi et al. 2021; Ondieki et al. 2017; Ndungu et al. 2016). This is consistent with as shown in Panel B of Table 1, among the milk samples of 940 farmers from two dairy cooperatives in two different counties, 47% of them failed the minimally required standards on added water posted by the government (Kenya Bureau Of Standards (KeBS)[6]). Compared to added water, dairy farmers are doing better on butterfat and bacteria: 16% of them failed on butterfat, and 23% of them failed on the bacteria. Overall, 62% of farmers' milk fails at least one of the KeBS's requirements on water adulteration, butterfat, and bacteria. A likely reason for low milk quality is that individual farmers face weak incentives to produce high-quality milk.

In the Kenyan dairy (formal) value chain, farmers are selling their milk to dairy cooperatives. Dairy cooperatives hire milk transporters to collect milk from farmers. Milk is not tested at this point, however, beyond a simple acceptance test (organoleptic tests: sight, smell, or taste or some simplified version of density tests) due to the prohibitively high cost (both time and monetary) of comprehensive milk tests, and thus, farmers are paid based solely on quantity once they pass the acceptance tests. As Figure A1 shows, transporters usually aggregate the

[5]The number is slightly higher for rural households: 59 percent of rural households own at least one cow.

[6]The KeBS standards require milk to have less than 4% added water, at least 3.25% butterfat, and grade of 2 or above for bacteria. Milk that fails any of three quality standards is illegal to trade.

milk from multiple farmers to fill larger milk cans, which are then transported to collection centers. The collection centers pour the milk together into cooling plants (usually containing 10 to 100 of these cans) for sale to processors. These processors then perform comprehensive milk quality tests. *Brookside Dairy Limited* (the biggest dominant milk processing company in Kenya, which controls 45 percent of the dairy market as of January 2016) is actively buying raw milk from cooperatives in the study region and always classifies milk into three categories based on the milk quality testing results: (1) accept the aggregated milk with a premium price; (2) accept it without a premium price; or (3) reject it.[7] The reason is that high-quality milk can be used to produce high-value products like buttermilk and yogurt. Medium-quality milk is limited to use in ultra-pasteurized liquid packets, but low-quality milk cannot be traded by law. If the milk is rejected, the entire vat of milk has to be destroyed by law, and the dairy cooperative will bear the loss. Therefore, quality incentives exist in the downstream markets but are not transmitted to the upstream farmers, which leads to a low-quality, low-price equilibrium.

2.2 Time-varying Milk Quality

To effectively address this market failure, it is essential first to determine whether quality defects are specific to individual farmers or vary over time. If milk quality is indeed farmer-specific, a single test could be sufficient to reveal relevant information. However, if milk quality fluctuates over time or is influenced by both farmer-specific and temporal factors, relying on a one-time test could result in both Type I and Type II errors.

First, I investigate the milk quality fluctuation at the aggregated can-level. Note that for the first round of quality monitoring at baseline, I tested the aggregated cans when they arrived at the cooperatives continuously for several days without interacting with individual farmers[8]. The two panels in Figure 1 display the average levels of added water and butterfat at the aggregated can-level for two counties, accompanied by 95% confidence intervals. The red dotted line represents the minimum national standards set by the KeBS, while the green dotted line indicates the processor's quality bonus thresholds. In the absence of the intervention, milk quality fluctuated over time but at different levels. If aggregated can samples are collected on Day 1 or Day 2 in Nyeri County, the one-time test results indicate that the average milk quality for added water meets the KeBS standard. However, if samples are taken between Day 4 and Day 13, the results show that the average milk quality for added

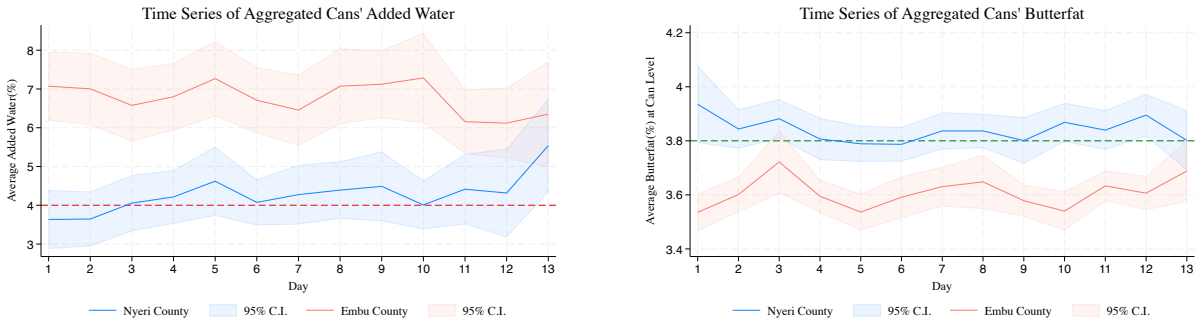
[7]The Brookside's quality bonus criteria is that the milk has less than 2% added water, more than 3.8% butterfat, and a grade of 5 or above for bacteria. The highest bonus cooperatives can get is 2Ksh, which is close to 6% of the base price of 45 Ksh on average.

[8]During the first round of quality monitoring at baseline, individual milk samples were collected from farmers exclusively on the final day of can-level testing. This was done to validate the model's predictions after getting the farmers' consent for milk testing. In subsequent monitoring rounds, however, milk samples were randomly collected throughout the monitoring periods rather than being restricted to the final day.

water fails to meet the KeBS standard. On Days 12 and 13, the 95% confidence intervals for the two counties overlap, suggesting that a random sample taken in both counties could lead to the conclusion that Embu County has less added water than Nyeri County. Nonetheless, Embu County demonstrates consistently better performance on most days compared to Nyeri County.

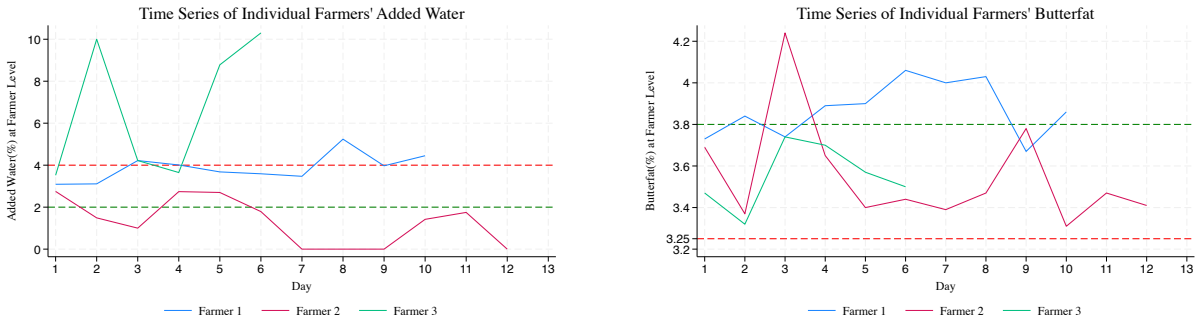
Second, the milk quality shows greater temporal variability at the individual farmer level. It is not an easy task to examine the milk quality at the farmer level since the action of taking farmers' milk samples continuously for several days will likely change their behavior until the testing ceases. Conveniently, several farmers with large milk volumes can fill the 50-liter aggregated cans by themselves without mixing with anyone else. This creates an opportunity to observe their milk quality over time by testing the aggregated cans instead of their individual milk. I plot the milk quality time series of the farmers whose milk filled the aggregated cans alone and showed up for more than five days in the two panels of Figure 2.

Figure 1: Can level Milk Quality Fluctuation within the Same Monitoring Period



Notes: This figure illustrates the average aggregated can-level milk quality fluctuations over time with 95% confidence intervals during the first round of quality monitoring at baseline. During this round, aggregated cans were tested continuously over several days upon arrival at the cooperatives, with no direct interaction with individual farmers.

Figure 2: Farmer level Milk Quality Fluctuation within the Same Monitoring Period



Notes: This figure illustrates farmer-level milk quality fluctuations over time. This specifically highlights the milk quality of farmers whose milk exclusively filled 50-liter aggregated cans without mixing with others, enabling their quality to be monitored through can-level testing without direct engagement. Notably, milk quality shows greater temporal variation at the farmer level compared to the aggregated can level.

As shown in the time series plots in Figure 2, farmers' milk quality fluctuates over time but at different levels.[9] Farmer 2 is clearly a different type of farmer compared to Farmer 1 and Farmer 3 in terms of added water, for example. Therefore, there are both the farmer-fixed component and the time-varying component of the quality defects. It is important to note that if I take individual milk samples from Farmer 1 and Farmer 2 on Day 5, the one-time testing result will show both farmers are "normal" type[10] of added water; if I take individual milk samples from them on Day 10, the one-time test result will show Farmer 1 is "bad" type and Farmer 2 is "good" type of added water. That said, farmers' milk quality types are not set in stone, and one-time tests might not be sufficient to reveal the quality information for all days. One-time tests can reflect the perfect information only for the days when farmers' milk samples have been taken and tested, but not on other days. It fails to fully capture the dynamics of the milk quality.

2.3 Digital Traceability System

For this dynamic moral hazard problem, I established a traceability system to trace the milk throughout the value chain so it can provide a basis for continuous monitoring on a daily basis. As discussed above, when milk is aggregated by transporters, it is collected into cans, which include the milk from about 3-10 farmers. Farmers' milk volume and which cans they fill vary every day. To establish a milk traceability system, I first assign a unique identification number to each aggregated milk can, as is shown in Figure A2, and then link farmers' unique member IDs to specific milk cans for every day's milk delivery. Specifically, I give milk transporters basic smartphones with a digital app installed; then, as Figure A3 shows, the transporters use the digital app to record the date, time, farmers' member IDs, aggregated milk can IDs, the milk volume, transporter name, collection route, and cooling plant every time they collect milk from farmers. Cooperative staff help to check the data and upload it to the server when transporters drop the milk at the cooling plant so that I can get access to the data every day. At the end of the study, smartphones are given to the transporters as compensation for their time. This traceability system helps to record who contributes to each aggregated milk can and their corresponding contribution percentages, which are used in the Bayesian statistical models.

When collecting the milk, transporters pour the milk to the aggregated containers on the truck, as is shown in Figure A4. Transporters are advised to pour farmers' milk into different aggregated containers randomly, which adds more variation to the can level composition

[9] Added water is primarily a day-to-day choice (time-varying); butterfat is determined by breed type (farmer-specific in the short-run) and animal feed (time-varying), and butterfat could also be affected by added water (time-varying).

[10] I define milk which meets the national standards but fails the processor's bonus standards as "normal" type; milk fails the national standards as "bad" type; and milk meets the processor's bonus standards as "good" type.

and improves the model performance. Specifically, before departing for milk collection, the aggregated containers are arranged sequentially on the truck according to their can IDs. In the field, milk from the first farmer is poured into the first can, the second farmer's milk into the second can, and so on. Once all the aggregated cans have been filled with milk from at least one farmer, a new round of pouring begins. During this round, transporters randomly select a starting can and then proceed sequentially through the containers. This process ensures that the neighboring farmers do not consistently mix their milk (e.g., the 1st and 2nd farmers), and it also prevents cases where the 1st and 23rd (if there are 22 cans in total) farmers' milk is always mixed in the same container.

3 A Bayesian Model of Quality Estimation

In addition to the traceability data, I tested these aggregated cans when they arrived at the cooling plant for several days.[11] As Figure A5 shows, before collecting samples from the aggregated cans, the field team used a plunger to thoroughly mix the milk by stirring it from top to bottom until it was well combined.[12] Then the field team pours the milk into small sample tubes with unique sample IDs and stores them with ice bricks to keep the milk fresh, while shipping to the milk testing lab, as is shown in Figure A6. Then, I link these aggregated milk container level quality test results to each farmer so that I can generate the inputs for model prediction. To be specific, I have a matrix of each farmer's contribution percentage for each aggregated can, and the corresponding aggregated can test results on different days.

Transporters are advised to pour farmers' milk into different aggregated containers randomly, which adds more variations to the can level composition. The model utilizes the variation from the (1) farmers' milk volume; (2) which aggregated cans they fill; (3) which other farmers are contributing to that can; and (4) corresponding aggregated can test results vary each day. The theoretical aggregated can quality should be the weighted average of the individual farmers' milk quality in the can.[13]

Compared to typical machine learning models or other methods[14], the main benefit of using the Bayesian model is that (1) it utilizes the prior knowledge that added water is a non-

[11] Until the number of can-level observations was 2 times the number of different farmers. This was to ensure that the model converges and performs well, as advised by results from a pilot study.

[12] This is proved to be necessary and useful during the pilot trial, and the field team includes this as one of the important steps into the milk sampling protocol.

[13] The weighted average relationship between can level quality and farmer level quality holds for added water and butterfat but not for bacteria. The added water and the butterfat are numerical data, while bacteria are ordinal data based on the testing techniques in this study, as is explained in Section 4.2.1.

[14] Fixed effect model for example.

negative value, and butterfat is a strictly positive value;[15] (2) it can directly incorporate the weighted average relationship between the aggregated can quality and individual farmers' milk quality into the model and associated likelihood functions. In addition, the Bayesian model also has two other clear benefits: (3) it performs well in different collection routes in different cooperatives, as it has less room for over-fitting (it does not rely on fine-tuning based on training samples, and the predictions are purely out-of-sample predictions); and (4) it provides not just the average milk quality (point estimate) for each farmer over the monitoring period but also a whole posterior distribution of each farmer's milk quality, which allows me to take into account the uncertainty and calculate the probability of farmers' milk quality type on any single day, which is important in this dynamic moral hazard problem, as is discussed in section 2.2.

3.1 Priors

To estimate the posterior distributions for individual milk quality, I need to give the model prior information about individual farmers' milk quality. Figure A7 shows the distributions of added water and butterfat at both the farmer level and the can level. I chose the prior distribution of farmer-level added water to be Half-Normal since added water is a non-negative value, and Panel A of Figure A7 shows that it is likely to be a Half-Normal distribution; I chose the prior distribution of farmer-level butterfat to be Gamma since butterfat is a strictly positive value, and Panel B of Figure A7 shows that it is likely to be a Gamma distribution. I chose the weakly informative priors for the Half-Normal and Gamma distribution parameters to let them be drawn from an Inverse Gamma distribution. All the details are included in the Technical Appendix C.

3.2 Likelihood

The theoretical aggregated can-level quality should be the weighted average of the individual farmers' milk quality in the can. The observed aggregated can-level quality typically has some measurement error around its theoretical mean, so I model the can-level quality to follow a Truncated Normal distribution[16], with the mean equal to the weighted average of the individual farmers' milk quality in the can. The likelihood is calculated based on the pdf of the Truncated Normal distribution, and the intuition is that the further away it is from the theoretical mean, the less likely it is. All additional details are included in the Technical Appendix C.

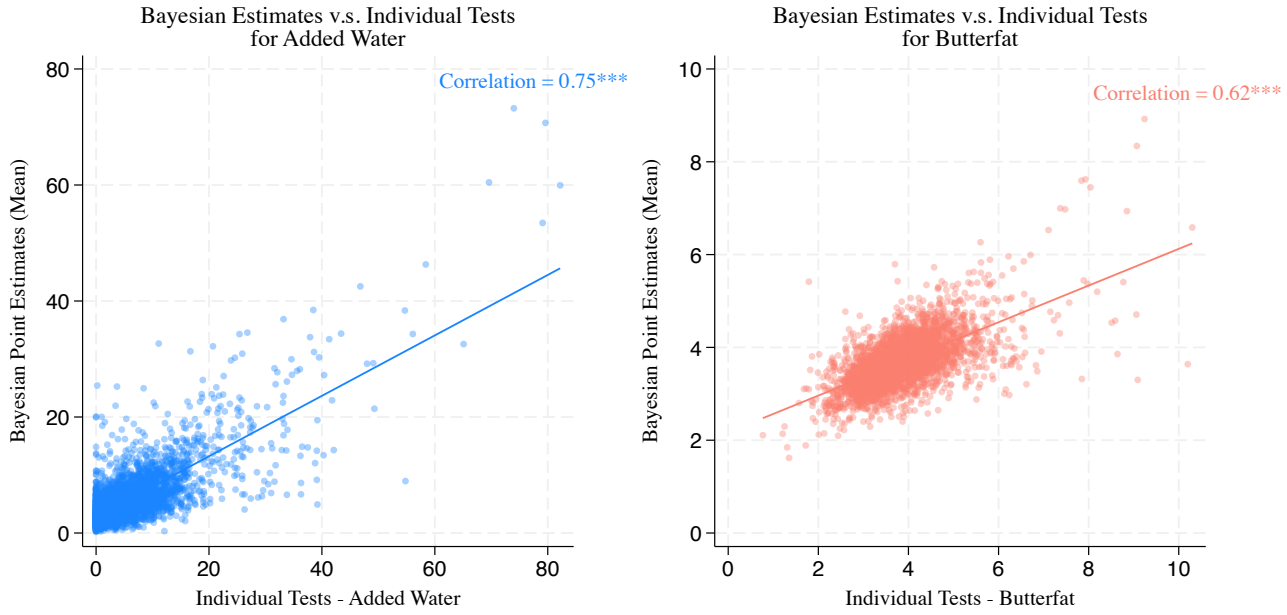
[15] In the next section, I show that I also have prior knowledge that individual farmers' added water is likely to follow a half-normal distribution, and butterfat is likely to follow a gamma distribution based on the farmer level histogram.

[16] Truncated at 0 since the quality parameters are non-negative values.

3.3 Posteriors and Model Performance

The Bayesian statistical models give the posterior point estimates of the farmers' quality parameters, which are the average quality over the monitoring period. Figure 3 compares the model point estimates to the one-time individual tests: the model-predicted quality shows a high correlation (at 1% level) with the individual milk test (0.75 for added water and 0.62 for butterfat) among 940 farmers.[17]

Figure 3: Comparison between Model Prediction and Individual Tests



Notes: This figure compares Bayesian point estimates of average milk quality over the monitoring period with random individual test results. The left panel focuses on added water, while the right panel highlights butterfat. Both the X-axis and Y-axis represent absolute quality levels measured as percentages. The Pearson correlation coefficients between the Bayesian point estimates and the individual test results are calculated, pooling data across all four rounds of quality monitoring. ***, **, and * represent significance at 1%, 5%, and 10% respectively.

In addition, the Bayesian statistical models give two other outcome variables that are also used for the analysis: First, the probability of each parameter failing the quality standards (based on the entire posterior distribution); which can be either interpreted as the chance of failing the standards if the milk is sampled and tested in any given day, or percentage of days that the farmers' milk fails the standards over the monitoring period. Second, I also classify farmers' milk quality as "bad", "normal", or "good," based on whichever category has the highest probability over the monitoring periods. Although I have discussed in section 2.2 that the farmer quality type is not fixed for every day, this classification is mainly meant to classify farmers' milk in a general sense during the monitoring period so that I can do

[17] I include additional ways of measuring model performance in the Technical Appendix C.

subgroup analysis based on the model-defined bad quality later on. It is also worth noticing that the model captured a larger percentage of farmers who were bad on quality.[18] I will come back to this difference when discussing the subgroup analysis using different definitions of bad quality, either by model or individual tests.

Lastly, there is a chance for prediction errors. Therefore, I explained the model results multiple times to the cooperatives' management until they fully digested the information, and both of the cooperatives recognized that the quality signals were useful and informative but could also be noisy. It is important to note that the prediction error for each round is independent (no systematic bias towards single farmers), as farmers' milk is poured into different cans randomly, so the chance of the same farmer being wrongly classified for two or more rounds is low.

4 Experimental Design

4.1 Experimental Context and Design

As mentioned in previous sections, the cooperatives are not able to test every farmer every day due to the prohibitively high cost (both time and monetary) of comprehensive milk tests. The model is developed to extract individual farmers' quality information from the pooled can-level tests, and thus, it could save testing costs. Besides this approach, the cooperatives could also consider conducting random testing, given the limited monitoring resources. Random monitoring gives perfect information for the days when milk has been sampled and tested but not other days. The cooperative updates its belief about how many days the farmer produces high- or low-quality milk. In contrast, the Bayesian models on aggregated can-level do not give perfect information for any single day but provide information about overall performance, including the average milk quality and the probability of producing high- or low-quality milk on each day, as derived from the entire posterior distribution discussed in section 3.3.

To test both of them in this dynamic moral hazard setting, I conducted an individual-level RCT to randomly assign 940 farmers from two different cooperatives to one of the three treatment conditions stratified by the milk collection routes. Group 1 farmers' model predicted quality information is shared with the cooperatives (as well as with the farmers themselves) as the reviews for the entire monitoring period. Group 2 farmers' one-time quality test information is shared with the cooperatives (as well as with the farmers themselves) as random

[18] As is shown in Panel B of Table 1, fifty-seven percent of farmers were classified by the model as bad quality on added water, while forty-seven percent of farmers were classified by individual tests as bad quality on added water; nineteen percent of farmers were classified by the model as bad quality on butterfat, while fourteen percent of farmers were classified by individual tests as bad quality on butterfat.

monitoring. Group 3 serves as the control for the evaluation.

In the model group, I tested the aggregated cans for several days (10 on average, depending on the size of collection routes) until the number of can-level observations was two times the number of different farmers. Then I shared the three outcomes (generated by the model as discussed in section 3.3) to both the farmers and the cooperatives: (1)average quality (from point estimates), (2)the probability of producing good-, normal-, or bad-quality for any given day (from posterior distributions), and (3) the general classification of quality types based on whichever category has the highest probability over the monitoring periods. The quality parameters shared in the model group include added water and butterfat, not bacteria, as the model's likelihood relies on the weighted average relationship between can-level quality and farmer-level quality, as mentioned in section 3.2. The weighted average relationship holds for added water and butterfat but not for bacteria. The added water and the butterfat are numerical data, while bacteria are ordinal data based on the testing techniques in this study. I will provide more information about the testing techniques in section 4.2.

In the random monitoring group, I informed the farmers that there would be a 10 percent chance of testing milk every day, and the field team took farmer samples once in the 10-day monitoring period. The individual samples were tested for added water, butterfat, and bacteria. I shared the information on all three parameters[19] with both the farmers and the cooperatives.

One important difference when sharing the information with farmers in these two treatment groups is that, in the random individual tests group, the field team informs the farmers of the testing result and its corresponding quality category of bad, normal, or good. In contrast, in the model group, the field team informs the farmers of the probability of being bad quality type even if it is a 5% chance. This has implications for the effects of information sharing since all farmers have a non-zero probability of being bad. I will revisit this in section 5.3.

I examine how cooperatives and farmers respond to this information. I have continuously shared the information for 3 rounds so that I can monitor the dynamic changes in the cooperatives' and farmers' behavior and milk quality. After I share the quality information from one round and before I start a new round of quality monitoring, there are one to two weeks off for the cooperatives and farmers to react. The first round of quality monitoring occurred for the cooperative in Nyeri in May 2024 and in June 2024 for the cooperative in Embu. Starting in July, I monitored quality in parallel on both cooperatives: the second round of quality monitoring was conducted in July, the third round of quality monitoring

[19]Although in the model group, I did not share the bacteria information as the model does not work for bacteria due to the data nature, I think it is still important and interesting to examine the cooperatives' reactions to the bacteria information. Later on, I observed that cooperatives respond differently and separately to each quality parameter.

was conducted in August, and the fourth round was in September. Note that I need to use the current round of quality monitoring results to evaluate the effects of the previous round of quality monitoring and information sharing. Figure A8 presents the experimental design and the timeline of this study.

4.2 Data

My main outcome data comes from two primary sources: milk quality data from the milk tests and farmers' behavior change from the survey data. In addition, I have the administrative data on milk volume and the attendance sheets for the cooperatives' training events. Cooperatives offer loans to the members through the Savings and Credit Cooperative Society (SACCO), from which I have also obtained farmers' credit data.

4.2.1 Milk Quality Data

This study focuses on three parameters of milk quality: added water, butterfat, and bacteria, following the processor's quality-price schemes. Among the three quality parameters, added water and butterfat are tested by milk analyzers[20] (see the Figure A9), which typically takes 1 minute per test.[21] The bacteria is tested using the Resazurin Test (RT), in which a prepared solution of resazurin dye is added to the milk sample. The color change after incubation indicates the milk quality: blue for good quality, purple or pink for normal quality, and colorless for bad quality.[22] Bacterial levels are determined by comparing the color to a reference chart, as is shown in Figure A10.

Note that I tested every aggregated can and also recorded traceability data for each can, so I have the model prediction results for all farmers in the study sample, regardless of the treatment arms. At the same time, I also have individual test results for all farmers in the study sample, regardless of the treatment arms, since we need to validate the model performance for each round. The only difference is that there are no model-predicted bacteria, but there are bacteria results from individual tests.

4.2.2 Surveys and Cooperatives' Administrative Data

I conducted baseline farmer surveys in May and June 2024 and endline farmer surveys in October 2024. The surveys capture farmers' key farming practices related to milk quality, including cow types and animal feeds for butterfat, the milking hygiene indicators, and

[20] I use the Ekomilk ULTRA PRO ultrasonic milk analyzer, produced by *Eon Trading INC* in Bulgaria.

[21] The lab staff usually tests 3 times per sample and records the last testing results to minimize the impact of the residual milk from the previous samples.

[22] In the RT test, 1 mL of 0.005% resazurin dye solution is added to 10 mL of the milk sample in a test tube, then incubated in a water bath at 37°C for 10 minutes. The test results are interpreted based on how quickly the color shifts, reflecting bacterial activity.

milk-storing techniques for bacteria. I have access to the cooperatives' records of farmers' daily milk delivery, which are the basis for milk payment, the attendance sheets for the cooperatives' training event, and also the farmers' credit limit data.

4.3 Summary Statistics and Randomization Check

Table 1 presents the summary statistics and a check of the randomization balance for a selected set of indicators. For each variable, I show the control mean in Column 1 and the difference between each treatment group and the control group in Column 2-3. Twenty percent of household heads are female, and the average respondent is 55 years old, has 10 years of education, has 4 household members, and owns 2.7 cattle; of those, 1.49 are lactating cows. On average, households have been keeping cattle for 18 years and have a daily production of 15 kgs of milk; of those, 11 kgs are sold to the cooperatives. Only 1 percent of households reported actively selling to informal traders. The price dairy farmers receive from selling milk to cooperatives is 41.3 ksh per kg, and they are paid on a monthly basis. At baseline, twenty-three percent of the households reported that they had been trained on butterfat, and twenty-five percent of the households reported that they had been trained on bacteria in the past 6 months. As a result, the milk quality tests show that the average added water in the farmers' milk is 6 percent, and forty-seven percent of the farmers will be classified as having bad quality added water based on individual tests, according to the national standards (KeBS); the model detected a higher percentage of farmers to be bad quality of added water, which is fifty-seven percent. The individual tests show that the average butterfat is 4.03%, and only fourteen percent of farmers were classified as bad quality according to the national standards (KeBS); the model detected a higher percentage of farmers to be bad quality of butterfat, which is nineteen percent. In addition, the individual tests show twenty-three percent of farmers had a bad quality of milk in terms of bacteria.

Overall, randomization appears successful - of the 40 regression coefficients in the table, only 2 are significant at the 10 percent level (household size and the number of years keeping cattle). Most importantly, the milk quality is balanced at baseline for all of the added water, butterfat, and bacteria. Further, there is no obvious channel in which the household size and the number of years of keeping cattle can directly affect the milk quality, but I still control both of these covariates in all regressions. In addition, all regressions throughout the paper were pre-specified as ANCOVA, and I control for baseline measures of outcomes.

Table 1: Baseline Summary Statistics and Experimental Balance

	(1) Mean (SD)	(2) Coefficient on Difference (Treatment - Control)	(3)
	Control	Random Tests	Model Detection
Panel A. Demographics			
Age	54.65	-0.66	-0.35
	(13.42)	(0.98)	(0.98)
=1 if female	0.19	0.01	0.02
		(0.03)	(0.03)
Years of education	9.90	-0.09	0.16
	(3.40)	(0.22)	(0.22)
Number of household members	4.07	0.12	-0.19*
	(1.49)	(0.11)	(0.11)
Total cattle owned	2.73	-0.02	0.02
	(1.84)	(0.13)	(0.14)
Number of lactating cows owned	1.49	0.02	-0.03
	(0.91)	(0.06)	(0.06)
Number of years keeping cattle	18.19	-1.82*	0.60
	(13.24)	(0.95)	(0.95)
Daily milk production (Kgs)	14.59	-0.03	-0.31
	(11.20)	(0.85)	(0.85)
Daily sales to cooperatives (Kgs)	11.16	0.08	-0.61
	(9.82)	(0.72)	(0.73)
=1 if selling to informal traders actively	0.01	-0.00	-0.00
		(0.01)	(0.01)
Payment frequency from cooperatives (Times per month)	1.00	0.00	0.00
	(0.00)	(0.01)	(0.01)
=1 if trained on butterfat	0.23	-0.00	0.00
	(0.42)	(0.03)	(0.03)
=1 if trained on bacteria	0.25	-0.01	0.00
	(0.43)	(0.03)	(0.03)
Panel B. Milk Quality			
Added water(%)	5.98	0.45	-0.49
	(7.03)	(0.50)	(0.49)
=1 if bad quality of added water (individual sample)	0.47	-0.00	0.01
		(0.03)	(0.03)
=1 if bad quality of added water (model detection by can sample)	0.57	0.02	-0.00
		(0.03)	(0.03)
Butterfat(%)	4.03	-0.02	-0.03
	(0.84)	(0.06)	(0.06)
=1 if bad quality of butterfat (individual sample)	0.14	0.03	0.01
		(0.03)	(0.03)
=1 if bad quality of butterfat (model detection by can sample)	0.19	-0.01	-0.01
		(0.03)	(0.03)
=1 if bad quality of bacteria (individual sample)	0.23	0.04	-0.04
		(0.03)	(0.03)

Notes: N=719 for Panel A, the variables in Panel A are from the baseline farmer surveys; N=940 for Panel B, the variables in Panel B are from the milk tests at the individual farmer level or the model detection based on the aggregated can sample tests. Bad quality refers to failing to meet the national KeBS standards for the corresponding quality parameter. The dependent variables are in rows; each row shows a coefficient from a separate regression on the respective dependent variable. Standard errors are shown in parentheses under the coefficients. The strata (milk collection route) fixed effect is included in all regressions. ***, **, and * represent significance at 1%, 5%, and 10% respectively.

5 Results

5.1 Dynamic Treatment Effects

I estimate dynamic treatment effects using different rounds of quality monitoring information as follows:

$$Y_{ist} = \sum_{t} \beta_t \left(RandomTests_{is} \times D_t\right) + \sum_{t} \gamma_t \left(Model_{is} \times D_t\right) + \eta Y_{is0} + \lambda_s + \phi_r + \theta X_i + \epsilon_{ist}, \quad (1)$$

where Y_{ist} is an outcome for individual i in strata s (collection route) at time t; t is defined as the quality monitoring rounds; $RandomTests_{is}$ is a treatment indicator that takes the value 1 if individual i is in the random test group and 0 otherwise; D_t is the binary variable equal to 1 for round t; $Model_{is}$ is a treatment indicator that takes the value 1 if individual i is in the model group and 0 otherwise; Y_{is0} is the baseline value of the outcome; λ_s is the collection route(strata) fixed effect; ϕ_r is the round fixed effect; X_i is a vector of baseline controls for family size and years of keeping cattle since treatment groups differ in those variables at baseline; and ϵ_{ist} is the error term. I cluster standard errors at the individual level, which is the level of randomization.

Figure 4 and Figure 5 plot the coefficients and confidence intervals from Equation (1) for both β_t (blue) and γ_t (red), which represent for the dynamic effects of random individual tests and model detection, respectively. Note that I use the individual test results as the outcomes to examine the treatment effects, and I need to use the current round of quality monitoring results to evaluate the effects of the previous round of quality monitoring and information sharing. Therefore, the quality monitoring rounds on the X-axis in each figure are from 2 to 4. The difference between Figure 4 and Figure 5 is that Figure 4 focuses on the change in absolute level of quality, which includes added water and butterfat, while Figure 5 focuses on the change in quality categories, which includes added water, butterfat, and bacteria.[23]

Figure 4a shows a continuous decrease in the added water level after each round of information sharing for both the random individual tests group and the model group, compared to the control group. The model group decreased more after 3 rounds of quality information sharing, and the effect of the model group is significant at the 5% level, while the effect of the random test group is insignificant. The magnitudes are sizeable for both groups: farmers in the random tests group have 1.01% less added water on average, and farmers in the model group have around 1.75% less added water on average. These translate to a 12.6 percent and a 21.9 percent decrease in added water compared to the baseline mean of the control

[23] I do not have an absolute level of bacteria because the testing techniques in this study only provide ordinal data, as discussed in Section 4.2.1.

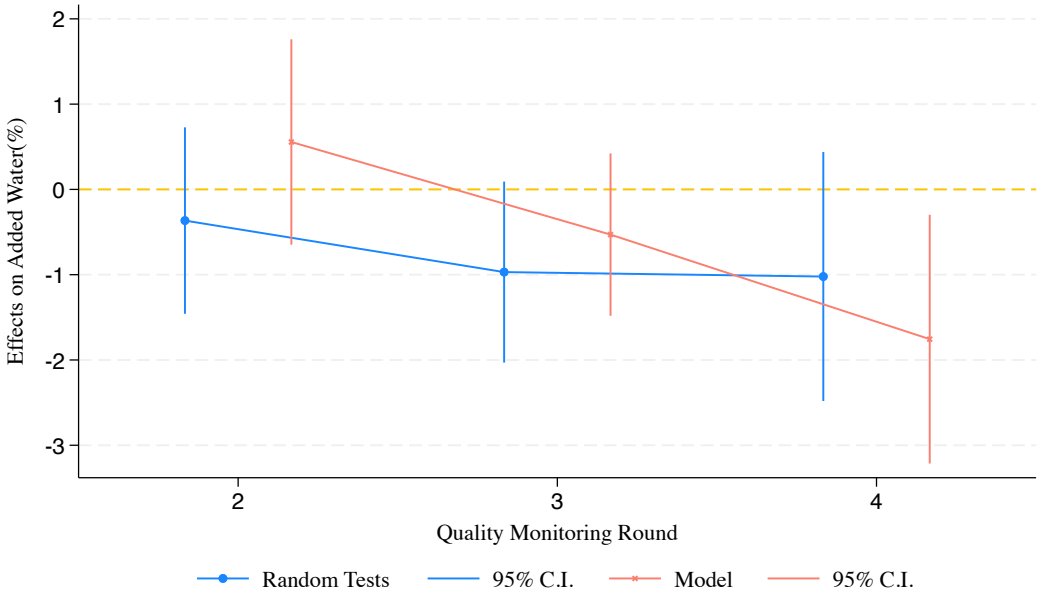
group (8% added water). The finding is consistent with my observation of the cooperatives' reactions to the information. New rounds of information update and reinforce the cooperatives' belief about farmers' quality, and cooperatives take action accordingly to regulate water adulteration.

Figure 4b shows that the butterfat increased after the information sharing (positive coefficients), but in all rounds, it fell to be indistinguishable from zero. One reason is that the average butterfat was very high from the beginning of the project. As Table 1 shows, the baseline control mean for butterfat is 4.03%, which is higher than the processors' quality rewarding requirement of 3.8%. Therefore, the room for improvement is small.

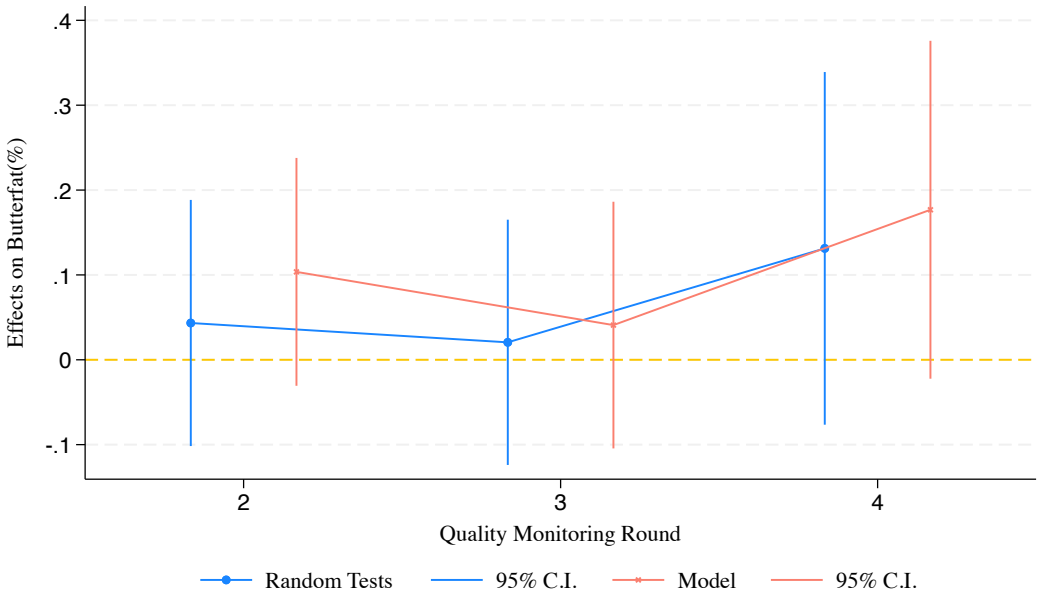
Figure 5 presents the effects on quality categories. Figure 5a shows that no significant amount of farmers changed their "bad" type on added water in the random tests group, while the share of farmers classified as "bad" type on added water reduced significantly in the model group. On butterfat, although both treatment groups did not significantly increase the level, the share of farmers classified as "bad" type on butterfat reduced significantly. There are no effects on the bacteria for farmers in the model group, as there is no information on bacteria shared in this group. The share of farmers who are classified as "bad" type of bacteria reduced only after 2nd round of information sharing, but the effects disappeared after the third round. It is also not surprising that I did not observe the cooperatives really emphasizing the bacteria; cooperatives trained the farmers with high bacterial load, but there were no real incentives for farmers to keep the quality standards on bacteria.

Figure 4: Dynamic Effects on Quality Level (%)

(a) Added Water



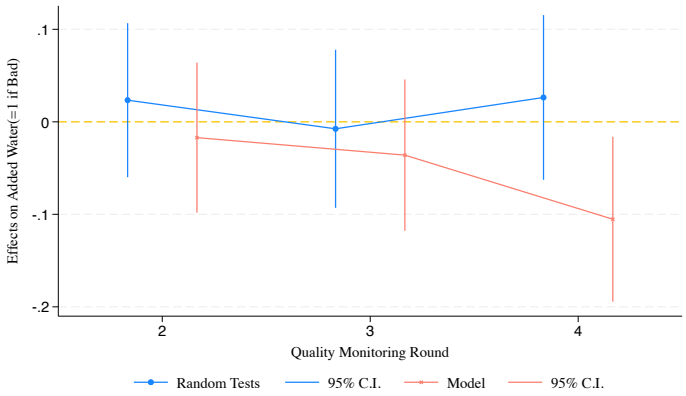
(b) Butterfat



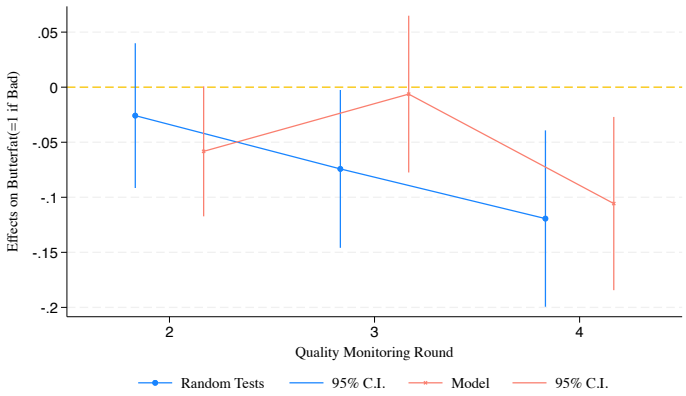
Notes: This figure presents the coefficients from the regression equation (1). The regressions include the baseline measurement of the outcome, baseline controls for the number of household members and the number of years keeping cattle, monitoring round fixed effects, and strata (collection route) fixed effects. Standard errors are clustered at the individual level, which is the level of randomization. The outcomes shown in this figure represent absolute quality levels measured as percentages based on the individual-level random tests. Bacterial data is excluded from the figure, as it is ordinal in nature, as explained in Section 4.2.1.

Figure 5: Dynamic Effects on Share of Farmers Classified as Bad Quality (=1 if Bad)

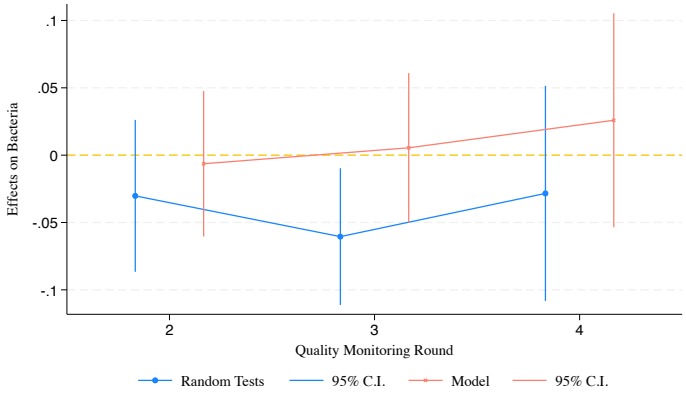
(a) Added Water



(b) Butterfat



(c) Bacteria



Notes: This figure presents the coefficients from the regression equation (1). The regressions include the baseline measurement of the outcome, baseline controls for the number of household members and the number of years keeping cattle, monitoring round fixed effects, and strata (collection route) fixed effects. Standard errors are clustered at the individual level, which is the level of randomization. The outcomes in this figure represent quality classifications based on a comparison of individual-level random test results against national KeBS standards. "Bad quality" indicates the failure to meet the KeBS standards for the corresponding quality parameter. Note that there is no bacteria information shared for model group farmers since the model works for numerical values, while bacterial data is ordinal in nature, as explained in Section 4.2.1. Therefore, the model group is expected to show no significant effect on bacteria in Figure 5c.

5.2 Effects on Milk Quality at Endline

I use the last round of quality monitoring as endline quality to examine the effects after 3 rounds of information sharing. To do it, I use a similar specification:

$$Y_{is} = \beta \text{RandomTests}_{is} + \gamma \text{Model}_{is} + \eta Y_{is0} + \lambda_s + \theta X_i + \varepsilon_{is}, \tag{2}$$

where Y_{is} is the value of the last round of quality monitoring results. Table 2 presents the effects of three rounds of information sharing on added water in Columns 1-2, the effects on butterfat in Columns 3-4, and the effect on bacteria in Column 5. Column 1 and Column 3 are the effects on the absolute quality level. Column 2, Column 4, and Column 5 are the effects on quality categories, defined by failing the minimal requirement of the national standards KeBS. I also provide a test for whether $\beta = \gamma$ at the bottom of the table.

I confirmed again that farmers in the model group have 1.75% less added water, which is a 21.9 percent reduction compared to the control group; farmers in the random tests group have 1.01% less added water, which is a 12.6 percent reduction compared to the control group. There are no significant effects on the level of butterfat, but the share of "bad" type butterfat farmers significantly reduced in both the random tests group and the model group. There are 11% and 10% fewer farmers classified as "bad" type butterfat for the random tests group and model group correspondingly, compared to the control group. Column 5 shows that there is little effect on the share of bad bacteria farmers in either group.

The bottom of the table shows tests for whether the effects are equal between the random tests group and the model group. Of the five outcomes, I only reject the equality for the bad water type category. Note that the model is able to capture a larger portion of "bad" type farmers (47% bad water farmers by individual tests, and 57% bad water farmers by model, as is shown in Table 1 Panel B.). This might also be the reason why the added water level is reduced more compared to the random tests group.

Table 2: Treatment Effects on Milk Quality at Endline

	Added Water(%)	=1 if Bad Water	Butterfat(%)	=1 if Bad Butterfat	=1 if Bad Bacteria
	(1)	(2)	(3)	(4)	(5)
Random Tests (β)	-1.01	0.02	0.10	-0.11***	-0.02
	(0.73)	(0.04)	(0.09)	(0.04)	(0.04)
Model Detection (γ)	-1.75**	-0.10**	0.13	-0.10**	-0.01
	(0.72)	(0.04)	(0.09)	(0.04)	(0.04)
p-value:					
$\beta = \gamma$	0.303	0.003	0.705	0.838	0.817
Observations	649	649	649	649	649
Control Mean of Depvar	8.00	0.65	3.86	0.28	0.21
Control SD of Depvar	9.57	0.48	1.05	0.45	0.41

Notes: This table presents results from running equation (2). Regressions include baseline measurements of outcome, collection route fixed effects, and baseline controls for the number of household members and the number of years keeping cattle. Column (1) and column (3) present the effect on absolute quality levels measured as percentages based on the individual-level random tests. Column (2), column (4), and column (5) present the effect on whether they were classified as "Bad Quality". "Bad quality" indicates the failure to meet the KeBS standards for the corresponding quality parameter. "Bad Quality" classifications in the dependent variables are based on a comparison of individual-level random test results against national KeBS standards. Note that there is no bacteria information shared for model group farmers since the model works for numerical values, while bacterial data is ordinal in nature, as explained in Section 4.2.1. Therefore, the model group is expected to show no significant effect on bacteria in column (5). ***, **, and * represent significance at 1%, 5%, and 10% respectively.

5.3 Subgroup Analysis

To further explore which group of farmers is driving the results, the following specification is used to conduct the subgroup analysis:

$$\begin{aligned} Y_{is} =& \beta(\text{RandomTests}_{is} \times \text{BadQuality}_{i0}) + \gamma(\text{Model}_{is} \times \text{BadQuality}_{i0}) \\ &+ \delta\text{RandomTests}_{is} + \mu\text{Model}_{is} + \alpha\text{BadQuality}_{i0} + \eta Y_{is0} + \lambda_s + \theta X_i + \varepsilon_{is}, \end{aligned} \tag{3}$$

where BadQuality_{i0} is the baseline quality indicator that takes the value 1 if individual i is classified as bad quality and 0 otherwise. Note that the bad quality can be defined by either the model or individual tests, and the model defined a larger portion of farmers as the bad type. I will use both bad-quality assignments to test for differences by subgroup. I will focus on the subgroup analysis for the three outcomes I find significantly improved in Table 2: Added water level, added water category, and butterfat category.

Table 3 Panel A presents the results when bad quality is defined by individual tests: It appears that bad-quality farmers on added water are the ones who drive the reduction in added water level for the random tests group. Note that the coefficient of the interaction term with the model (γ) is not significant, mainly because the "good" farmers on added water (defined by individual tests) also contribute to the reduction of the average added water level in the model group. The coefficient of the interaction term with the model (β) is very close to being significant at 10% level. Both "good" farmers and "bad" farmers on added

water have reduced the chance of being bad on added water. It appears that if I define the bad quality of butterfat by individual tests, the "good" farmers defined by individual tests at baseline are the ones who drive the results of reduction in the share of bad butterfat farmers.

Table 3 Panel B presents the results when bad quality is defined by the model predictions: the results are clearly mainly driven by the "bad" quality farmers at baseline for all three outcomes of added water level, added water category, and butterfat category. It is worth noticing that "bad" added water farmers reduced the chance of being classified as bad added water type, but the "good" added water farmers increased the chance of being classified as bad added water type. This explains why there is no significant effect on the share of "bad" type added water farmers in the pooled regressions in Table 2 for the random tests group.

Overall, both "good" and "bad" farmers, as defined by the model at baseline, contribute to the improvement of milk quality. The intuition for this is that, as described in section 4.1, the field team informs the farmers of the probability of being bad quality type on any single day, no matter how small it is, even if it is a 5% chance, while in the random individual tests group, the field team informs them of their testing results and whether it is good or bad. Therefore, this potentially causes the difference between the behavior of "good" farmers in the model group and in the random monitoring group since "good" farmers are assigned a non-zero probability of being bad, and "bad" farmers are simply informed that they were not bad upon testing. In addition, the model defined a larger portion of farmers as bad type, and the model defined bad quality type farmers at baseline were the ones who drive the results.

I plot the cumulative distribution function (CDF) in Figure A12 for added water and butterfat and the subgroup CDF for added water in Figure A12a and subgroup CDF for butterfat in Figure A12b. These plots further confirmed that "good" farmers at baseline behaved differently in the model group compared to the random tests group, as farmers in the model group feel being monitored constantly since no matter how small their chances of being classified as "bad" farmers, we notify them. In contrast, farmers in the random tests group were notified that their milk quality was good, if that is what the individual testing results showed. To be specific, in Panel (a) and Panel (b) of Figure A12a, no matter whether we defined "good" farmers based on the model or the individual tests, the "good" farmers in the random tests group were adding more water compared to the farmers in the control group. Only "bad" farmers in the random tests group were adding less water compared to the farmers in the control group, as is shown in Panel (c) and Panel (d) in Figure A12a. In contrast, both "good" farmers and "bad" farmers in the model group behaved better than the farmers of corresponding categories in the control group, as farmers in the model group felt being monitored constantly.

Table 3: Treatment Effects on Milk Quality at Endline (Subgroup by Bad Baseline Quality)

	Added Water(%)	=1 if Bad Water	Butterfat(%)	=1 if Bad Butterfat	=1 if Bad Bacteria
	(1)	(2)	(3)	(4)	(5)
Panel A: Bad Quality by Individual Tests					
Random Tests × Bad Quality (β)	-2.26	-0.07	-0.21	-0.08	-0.02
	(1.46)	(0.09)	(0.26)	(0.11)	(0.08)
Random Tests	0.05	0.06	0.14	-0.09**	-0.01
	(1.00)	(0.06)	(0.10)	(0.04)	(0.04)
Model Detection × Bad Quality (γ)	-1.21	0.01	-0.32	-0.03	-0.01
	(1.44)	(0.09)	(0.26)	(0.11)	(0.09)
Model Detection	-1.17	-0.11*	0.18*	-0.09**	-0.01
	(0.99)	(0.06)	(0.10)	(0.04)	(0.04)
=1 if Bad Quality by Individual Tests	0.26	0.32***	0.01	0.29***	0.01
	(1.17)	(0.06)	(0.21)	(0.08)	(0.06)
p-value:					
$\beta = \gamma$	0.462	0.339	0.640	0.622	0.896
Observations	649	649	649	649	649
Control × Good Baseline Quality : Mean	5.59	0.50	3.91	0.24	0.23
Control × Good Baseline Quality : SD	8.09	0.50	0.98	0.43	0.42
Panel B: Bad Quality by Model					
Random Tests × Bad Quality (β)	-2.61*	-0.17*	0.25	-0.25**	
	(1.47)	(0.09)	(0.25)	(0.10)	
Random Tests	0.43	0.12*	0.05	-0.05	
	(1.09)	(0.07)	(0.10)	(0.04)	
Model Detection × Bad Quality (γ)	-1.41	-0.13	0.15	-0.17*	
	(1.46)	(0.09)	(0.24)	(0.10)	
Model Detection	-0.98	-0.03	0.10	-0.06	
	(1.09)	(0.07)	(0.10)	(0.04)	
=1 if Bad Quality by Model	1.44	0.19***	-0.26	0.33***	
	(1.11)	(0.07)	(0.18)	(0.07)	
p-value:					
$\beta = \gamma$	0.409	0.630	0.673	0.392	
Observations	649	649	649	649	
Control × Good Baseline Quality : Mean	5.42	0.46	3.95	0.20	
Control × Good Baseline Quality : SD	8.38	0.50	0.91	0.40	

Notes: This table presents results from running equation (3). Regressions include baseline measurements of outcome, collection route fixed effects, and baseline controls for the number of household members and the number of years keeping cattle. Column (1) and column (3) present the effect on absolute quality levels measured as percentages based on the individual-level random tests. Column (2), column (4), and column (5) present the effect on whether they were classified as "Bad Quality". "Bad quality" indicates the failure to meet the KeBS standards for the corresponding quality parameter. "Bad Quality" classifications in the dependent variables are determined by comparing individual-level random test results to national KeBS standards. For the regressors, "Bad Quality" classifications are derived from individual-level random tests in Panel A and from model detection in Panel B. Note that there is no bacteria information shared for model group farmers since the model works for numerical values, while bacterial data is ordinal in nature, as explained in Section 4.2.1. Therefore, the model group is expected to show no significant effect on bacteria in column (5) of Panel A. Additionally, no regression is conducted for column (5) in Panel B for the same reason. ***, **, and * represent significance at 1%, 5%, and 10% respectively.

5.4 Effects on Training

After I shared the quality information, cooperatives contacted the regional *Brookside* trainers[24] to help to train low-quality farmers targeting different quality parameters.

Table 4 presents the effects on training. Panel A shows that there is a significantly higher percentage of farmers in both the random tests group and model group who have been trained on butterfat, and only farmers in the random tests group have a significantly higher

[24]Figure A shows the training sessions.

percentage were trained on bacteria since I did not share the bacteria information for the model group. Panel B is the subgroup analysis using individual tests to define bad quality, and Panel C is the subgroup analysis using the model to define bad quality. Both Panel B and C show that farmers who have bad quality at baseline are mainly the ones who got the training, although I do see some spillover of the training to farmers who have good quality of butterfat at the baseline when we use the model to define quality type in the random tests group.

The butterfat training covered the best animal feeding practices, and farmers' responses are presented in Table A2. Overall, there is no significant evidence that average treatment farmers used more concentrates and dry matter in animal feed to boost butterfat. One reason is that the average butterfat was very high from the beginning of the project. As Table 1 shows, the baseline control mean for butterfat is 4.03%, which is higher than the processors' quality rewarding requirement of 3.8%. Therefore, the room for improvement is small. It appears that farmers have already used an appropriate level of concentrates on average.[25]

The bacteria training covered recommended milking hygiene, and farmers' responses are presented in Table A3. In column (3), a significantly larger percentage (12 percent more) of farmers in the random individual tests group[26] used milking jelly after milking, and a significant 8 percent more farmers in the random tests group used different towels for milking different cows. Both of these practices are highly recommended by agronomists for milking hygiene. However, when it comes to practices that incur monetary costs, there is no evidence showing treatment farmers are following the recommended practices, which include using aluminum cans to store milk and not mixing evening milk and morning milk, then selling to the cooperatives. It is understandable that without clear quality rewards, farmers have low incentives to upgrade quality, especially when it is costly. In fact, there was a significant effect on reducing the probability of farmers failing the minimal required national standards of bacteria after the second round of information sharing, but the effects disappeared after the 3rd round. This suggests that cooperatives will need to provide clear quality incentives to farmers to have a long-lasting effect.

[25]Note that there is an optimal ratio of using concentrates and dry matter in the animal feed; if used excessively, it can also reduce the butterfat in the cow milk.

[26]Note that there is no bacteria information shared for farmers in the model group.

Table 4: Treatment Effects on Training

	=1 if Trained on Butterfat	=1 if Trained on Bacteria
	(1)	(2)
Panel A: Pooled Regression		
Random Tests (β)	0.11***	0.17***
	(0.02)	(0.02)
Model Detection (γ)	0.09***	-0.01
	(0.02)	(0.02)
p-value:		
$\beta = \gamma$	0.396	0.000
Observations	719	719
Control Mean of Depvar	0.18	0.21
Control SD of Depvar	0.38	0.41
Panel B: Interaction with Baseline Quality by Individual Tests		
Random Tests × Bad Quality (β)	0.50***	0.71***
	(0.05)	(0.03)
Random Tests	0.02	-0.01
	(0.02)	(0.02)
Model Detection × Bad Quality (γ)	0.48***	-0.01
	(0.05)	(0.04)
Model Detection	0.02	-0.01
	(0.02)	(0.02)
=1 if Bad Quality by Individual Tests	0.01	-0.00
	(0.04)	(0.03)
p-value:		
$\beta = \gamma$	0.650	0.000
Observations	719	719
Control × Good Baseline Quality : Mean	0.19	0.21
Control × Good Baseline Quality : SD	0.39	0.41
Panel C: Interaction with Baseline Quality by Model		
Random Tests × Bad Quality (β)	0.29***	
	(0.05)	
Random Tests	0.06**	
	(0.02)	
Model Detection × Bad Quality (γ)	0.46***	
	(0.05)	
Model Detection	0.02	
	(0.02)	
=1 if Bad Quality by Model	-0.00	
	(0.04)	
p-value:		
$\beta = \gamma$	0.001	
Observations	719	
Control × Good Baseline Quality : Mean	0.19	
Control × Good Baseline Quality : SD	0.39	

Notes: This table presents results from running equation (2) for Panel A, and equation (3) for Panel B and Panel C. Regressions include baseline measurements of outcome, collection route fixed effects, and baseline controls for the number of household members and the number of years keeping cattle. For the regressors, "Bad Quality" classifications are derived from individual-level random tests in Panel B and from model detection in Panel C. "Bad quality" indicates the failure to meet the KeBS standards for the corresponding quality parameter. Note that there is no bacteria information shared for model group farmers since the model works for numerical values, while bacterial data is ordinal in nature, as explained in Section 4.2.1. Therefore, the model group is expected to show no significant effect on bacteria in column (2) of both Panel A and Panel B. Additionally, no regression is conducted for column (2) in Panel C for the same reason. ***, **, and * represent significance at 1%, 5%, and 10% respectively.

5.5 Credit Access

I obtained administrative data from the SACCO regarding the credit limits of their members. As is shown in Table 5, the cooperatives increased the credit limit to those farmers who want to buy animal feeds and have never added water to the milk, defined by individual test in Panel B and by Model in Panel C, as the cooperatives believe that these farmers are more trustworthy and will be more likely to pay back the loans. It appears that the flow of quality information will also facilitate the flow of credit provision along the value chain.

Farmers in the model group who had never added water to the milk, defined by the model, enjoyed a 57-dollar higher credit limit compared to the farmers in the control group who had added water to the milk. The effect size is 42 dollars if we define never added water to the milk based on individual tests. The effect sizes are also sizable for farmers in the random tests group: 35 dollars if water is never added to milk, as defined by the model, and 28 dollars if water is never added to milk, as defined by the individual tests.

The increase in credit limits translated to the increase in purchasing animal feed on credit. As is shown in Column (2) of Table 5, Farmers in the model group who had never added water to the milk, defined by the model, purchased 30 dollars more animal feed on credit, compared to the farmers in the control group who had added water to the milk. The effect is 19 dollars if water is never added to milk, as defined by the individual tests, 26 dollars and 3 dollars for farmers in the random tests group if water is never added to milk, defined by the model, and by individual tests, respectively.

Note that outcomes in column (1) are derived from the cooperative's administrative data (SACCO), while outcomes in column (2) are based on the endline farmer survey. The survey data includes loans obtained through all channels, not limited to the SACCO, for the purchase of animal feeds. The number of observations differs between the two columns because the surveys were conducted on a subset of farmers from the study sample.

Table 5: Treatment Effects on Credit Access

	Credit Limit (USD) in the Current Month	Feed on Credit (USD) in the Past 30 days
	(1)	(2)
Panel A: Pooled Regression		
Random Tests (β)	-0.47	5.07
	(7.80)	(4.17)
Model Detection (γ)	8.32	6.31
	(7.63)	(4.17)
p-value:		
$\beta = \gamma$	0.251	0.766
Observations	402	314
Control Mean of Depvar	93.90	12.27
Control SD of Depvar	84.83	20.69
Panel B: Interaction with Never Added Water by Individual Tests		
Random Tests × Never Added Water (β)	28.32	3.40
	(21.04)	(12.30)
Random Tests	-4.97	4.53
	(8.54)	(4.43)
Model Detection × Never Added Water (γ)	42.04**	19.37*
	(20.36)	(11.62)
Model Detection	0.97	2.62
	(8.39)	(4.53)
=1 if Never Added Water by Individual Tests	-22.09	0.45
	(15.40)	(8.89)
p-value:		
$\beta = \gamma$	0.488	0.162
Observations	402	314
Control × Have Added Water : Mean	93.82	11.87
Control × Have Added Water : SD	82.55	19.74
Panel C: Interaction with Never Added Water by Model		
Random Tests × Never Added Water (β)	34.96	26.37*
	(27.36)	(15.54)
Random Tests	-4.43	2.55
	(8.25)	(4.33)
Model Detection × Never Added Water (γ)	56.99**	29.87**
	(27.53)	(14.93)
Model Detection	1.60	2.41
	(8.00)	(4.40)
=1 if Never Added Water by Model	-17.65	-10.97
	(23.04)	(12.87)
p-value:		
$\beta = \gamma$	0.301	0.764
Observations	402	314
Control × Have Added Water : Mean	92.22	12.31
Control × Have Added Water : SD	86.12	20.72

Notes: This table presents results from running equation (2) for Panel A, and equation (3) for Panel B and Panel C. Regressions include baseline measurements of outcome, collection route fixed effects, and baseline controls for the number of household members and the number of years keeping cattle. Outcomes in column (1) are derived from the cooperative's administrative data (SACCO), while outcomes in column (2) are based on the endline farmer survey. The survey data includes loans obtained through all channels, not limited to the SACCO, for the purchase of animal feeds. The number of observations differs between the two columns because the surveys were conducted on a subset of farmers from the study sample. For the regressors, "Never Added Water" classifications are derived from individual-level random tests in Panel B and from model detection in Panel C. "Never Added Water" indicates that the added water level consistently met KeBS standards across all four rounds of quality monitoring. ***, **, and * represent significance at 1%, 5%, and 10% respectively.

5.6 Cost Efficient Analysis

The results so far highlight the differences in milk quality improvements across the treatment groups. I calculate the improvements in added water and butterfat for different treatment arms across different monitoring periods, as shown in Figure A11.

Taking added water as an example, the average added water increased by 2.87 percentage points in the Control Group at the endline, by 1.59 percentage points in the Random Tests Group, and by only 0.16 percentage points in the Model Group. In terms of improvement, the Model Group demonstrates a significant impact, with an improvement of 2.71 percentage points (calculated by subtracting the baseline difference: 2.87 - 0.16), while the Random Tests Group shows an improvement of 1.28 percentage points (1.59 - 0.16). For the Model Group, can samples are tested until the number of can-level observations is twice the number of different farmers, resulting in a monetary cost that is double. However, the improvement is more than doubled, with the model achieving 2.12 (which is 2.71/1.28) times the improvement.

This back-of-the-envelope calculation indicates that achieved a more significant reduction in added water per testing dollar spent. Additionally, the model approach saves time on individual milk samples taken in the field, and it provides aggregated can-level quality information to help differentiate good milk from bad milk before it is mixed at the cooling plant. These advantages suggest that this new quality monitoring approach could be a preferable method for continuously monitoring and regulating milk quality.

6 Conclusion

This study aims to promote the transmission of quality incentives along the value chain from downstream buyers to upstream farmers, which is crucial for farmers to improve the quality of agricultural products and then transit towards market-driven, commercial agriculture. A longstanding barrier that has prevented the transmission of quality incentives in the dairy sector is the high testing cost of individual milk. My study addresses the high testing costs for individuals' milk through the digital traceability system and an innovative quality monitoring method based on Bayesian statistical models. Using pooled milk can testing combined with the traceability data; I developed a scalable solution that infers individual farmers' milk quality without incurring the full costs of testing each farmer separately. The model-predicted quality correlates significantly with individual test results and offers a reliable alternative, with a correlation of 0.75 for water adulteration and 0.62 for butterfat.

I also evaluate the effects of reducing quality information frictions by revealing randomly selected farmers' milk quality information either by the model or by random individual tests

to both the cooperatives and farmers. Experimental results highlight the effectiveness of this new monitoring approach: farmers in the model group reduced added water by a significant 21.9% relative to the control group, whereas those in the individual test group exhibited an insignificant reduction of 12.6%. The model outperforms conventional random testing across most metrics. A back-of-the-envelope calculation indicates that the model provides a cost-effective means of improving milk quality, making it a promising tool for continuous quality monitoring and scalable implementation.

I observed that cooperatives increased credit limits for farmers who consistently produce high-quality milk and seek loans to purchase animal feed. This decision reflects the cooperatives' belief that these farmers are more trustworthy and more likely to repay loans. This relationship suggests that the dissemination of quality information can facilitate credit allocation along the value chain. Enhanced access to credit for high-quality producers generates a positive feedback loop within the value chain, characterized by improved product quality, expanded credit opportunities, and increased income (Casaburi and Willis 2024). This paper contributes to the literature emphasizing the necessity of engaging multiple stakeholders across the value chain rather than focusing on isolated actors (Bellemare et al. 2022). The integration of a traceability system with Bayesian statistical models introduces novel data and methodological approaches, offering important implications for future analyses of value chain dynamics.

Compared to low-income countries, most developed economies have already required food firms to trace food, feed, and ingredients through all stages of production, processing, and distribution to the final consumer. Examples include but are not limited to regulations from *The General Food Law Regulation of the European Union* since 2005, and regulations from the *Food and Drug Administration (FDA)* in the U.S. since 2006. Kenya is currently developing new legislation to implement a quality-based milk payment system. Since October 2024, the *Kenya Cabinet Secretary for Agriculture* has collaborated with the *CGIAR International Livestock Research Institute (ILRI)* to shape this policy. My results speak directly to these attempts to drive the structural transformation of the Kenyan dairy sector. Beyond Kenya, the traceability system and the new quality monitoring approach have the potential to be scaled up to not only dairy cooperatives in many other countries but also a wide range of agricultural commodities.

References

Abate, Gashaw, Tanguy Bernard, Alain de Janvry, Elisabeth Sadoulet, and Carly Trachtman (2021). "Introducing quality certification in staple food markets in Sub-Saharan Africa: Four conditions for successful implementation". *Food Policy* 105: 102173.

Abate, Gashaw, Tanguy Bernard, Alain De Janvry, and Elisabeth Sadoulet (2022). "Quality-graded Wheat Value Chain Development and Agricultural Transformation in Ethiopia". *Unpublished*.

Aggarwal, Shilpa, Susan Godlonton, Ammar Kawash, Jonathan Robison, Alan Spearot, and Guanghong Xu (2024a). "The Value of Value Chains: An Experiment linking Farmers' Cooperatives with a Maize Processor in Rwanda". *Working Paper*.

Aggarwal, Shilpa, Athanase Nduwumuremyi, Jonathan Robison, Youngwoo Song, Alan Spearot, and Guanghong Xu (2024b). "Green Shoots: Seeding a Market for Certified Cassava Cuttings among Cooperatives in Rwanda". *Unpublished*.

Akerlof, George A. (1970). "The Market for "Lemons": Quality Uncertainty and the Market Mechanism". *The Quarterly Journal of Economics* 84 (3): 488–500.

Atkin, David, Amit K. Khandelwal, and Adam Osman (2017). "Exporting and Firm Performance: Evidence from a Randomized Experiment". *The Quarterly Journal of Economics* 132 (2): 551–615.

Bai, Jie (2024). "Melons as Lemons: Asymmetric Information, Consumer Learning and Seller Reputation". *Review of Economic Studies*.

Bai, Jie, Lauren Falcao Bergquist, Ameet Morjaria, Russell Morton, and Yulu Tang (2024). "Demand-side Incentives and Quality Upgrading in Uganda's Coffee Supply Chain". *Unpublished*.

Ball, Ian and Jan Knoepfle (2024). "Should the Timing of Inspections be Predictable?" *Working Paper*.

Bellemare, Marc F., Jeffrey R. Bloem, and Sunghun Lim (2022). "Chapter 89 - Producers, consumers, and value chains in low- and middle-income countries". *Handbook of Agricultural Economics*. Ed. by Christopher B. Barrett and David R. Just. Vol. 6. Elsevier: 4933–4996.

Bergquist, Lauren Falcao and Meredith Startz (2024). "Quality and Contracting in Honey Supply Chains in Ethiopia". *Unpublished*.

Bernard, Tanguy, Alain de Janvry, Samba Mbaye, and Elisabeth Sadoulet (2017). "Expected Product Market Reforms and Technology Adoption by Senegalese Onion Producers". *American Journal of Agricultural Economics* 99 (4): 1096–1115.

Bold, Tessa, Selene Ghisolfi, Frances Nsonzi, and Jakob Svensson (2022). "Market Access and Quality Upgrading: Evidence from Four Field Experiments". *American Economic Review* 112 (8): 2518–2552.

Casaburi, Lorenzo and Jack Willis (2024). "Value chain microfinance". *Oxford Review of Economic Policy* 40 (1): 160–175.

Deutschmann, Joshua W., Tanguy Bernard, and Ouambi Yameogo (2024). "Flexible contracting and quality upgrading under market volatility: experimental evidence from Senegal". *Working Paper*.

Do Nascimento Miguel, Jérémy (2024). "Returns to quality in rural agricultural markets: Evidence from wheat markets in Ethiopia". *Journal of Development Economics* 171: 103336.

Hansman, Christopher, Jonas Hjort, Gianmarco León-Ciliotta, and Matthieu Teachout (2020). "Vertical Integration, Supplier Behavior, and Quality Upgrading among Exporters". *Journal of Political Economy* 128 (9): 3570–3625.

Hoffmann, Vivian, Sarah Kariuki, Janneke Pieters, and Mark Treurniet (2023). "Upside risk, consumption value, and market returns to food safety". *American Journal of Agricultural Economics* 105 (3): 914–939.

Jin, Ginger Zhe and Phillip Leslie (2003). "The Effect of Information on Product Quality: Evidence from Restaurant Hygiene Grade Cards*". *The Quarterly Journal of Economics* 118 (2): 409–451.

Kelley, Erin M., Gregory Lane, and David Schönholzer (2024). "Monitoring in Small Firms: Experimental Evidence from Kenyan Public Transit". *American Economic Review* 114 (10): 3119–3160.

KIPPRA (2018). "Kenya Economic Report 2018". Tech. rep. Kenya Institute for Public Policy Research and Analysis (KIPPRA).

KNBS (2019). "Economic Survey 2019". Tech. rep. Kenya Bureau of Statistics.

Lane, Gregory, David Schönholzer, and Erin Kelley (2024). "Information and Strategy in Lemon Markets: Improving Safety in Informal Transit". *Working Paper*.

Macchiavello, Rocco and Josepa Miquel-Florensa (2019). "Buyer-Driven Upgrading in GVCs: The Sustainable Quality Program in Colombia". Tech. rep. 13935. C.E.P.R. Discussion Papers.

Magnan, Nicholas, Vivian Hoffmann, Nelson Opoku, Gissele Gajate Garrido, and Daniel Akwasi Kanyam (2021). "Information, technology, and market rewards: Incentivizing aflatoxin control in Ghana". *Journal of Development Economics* 151: 102620.

Ndungu, Teresiah W, Patrick S Muliro, Mary Omwamba, Gerard Oosterwijk, and Anton Jansen (2016). "Quality control of raw milk in the smallholder collection and bulking enterprises in Nakuru and Nyandarua Counties, Kenya". *African Journal of Food Science* 10 (5): 70–78.

Nyokabi, Simon N., Imke J.M. de Boer, Pieternel A. Luning, Luke Korir, Johanna Lindahl, Bernard Bett, and Simon J. Oosting (2021). "Milk quality along dairy farming systems and associated value chains in Kenya: An analysis of composition, contamination and adulteration". *Food Control* 119: 107482.

Ondieki, George Kiage, Jackson Nyarongi Ombui, Mark Obonyo, Zeinab Gura, Jane Githuku, Austine Bitek Orinde, and Joseph Kangangi Gikunju (2017). "Antimicrobial residues and compositional quality of informally marketed raw cow milk, Lamu West Sub-County, Kenya, 2015". *Pan African Medical Journal* 28.

Park, Sangyoon, Zhaoneng Yuan, and Hongsong Zhang (2023). "Technology Training, Buyer-Supplier Relationship, and Quality Upgrading in an Agricultural Supply Chain". *The Review of Economics and Statistics*: 1–46.

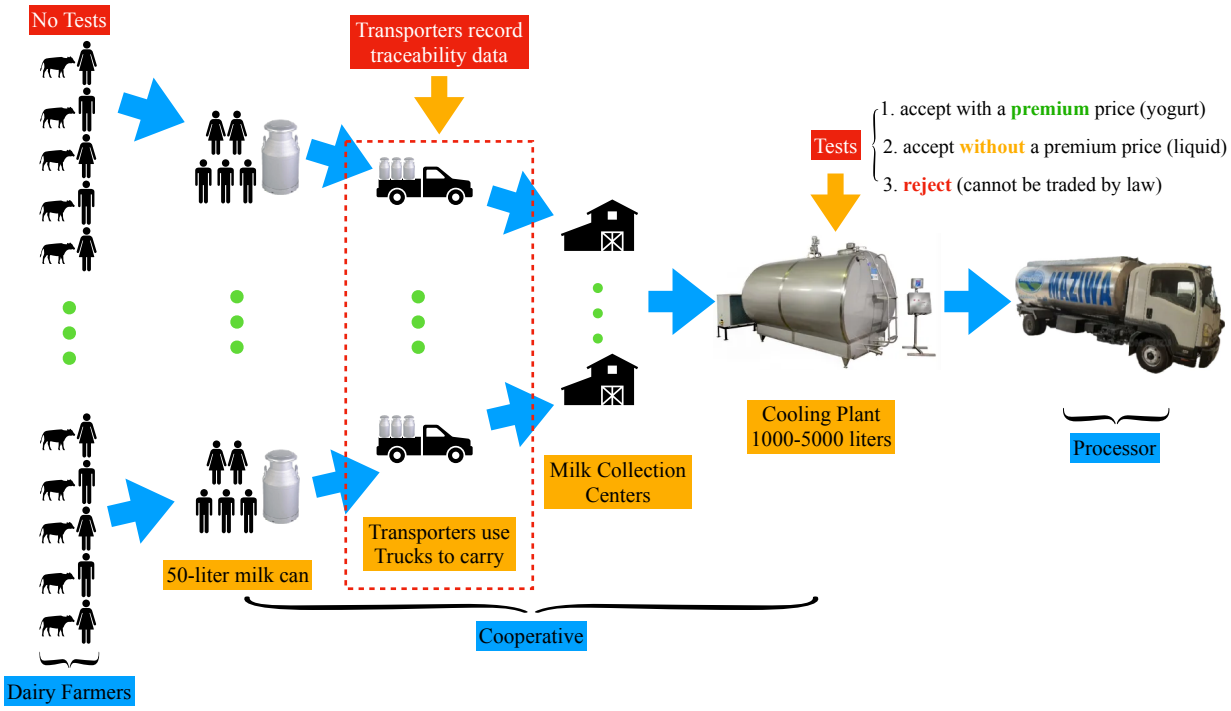
Rao, Manaswini and Ashish Shenoy (2023). "Got (Clean) Milk? Governance, Incentives, and Collective Action in Indian Dairy Cooperatives". *Journal of Economic Behavior & Organization* 212: 708–722.

Varas, Felipe, Iván Marinovic, and Andrzej Skrzypacz (2020). "Random Inspections and Periodic Reviews: Optimal Dynamic Monitoring". *The Review of Economic Studies* 87 (6): 2893–2937.

Viscusi, W. Kip (1978). “A Note on "Lemons" Markets with Quality Certification”. *The Bell Journal of Economics* 9 (1): 277–279.

A Appendix

Figure A1: Kenyan Dairy (Formal) Value Chain

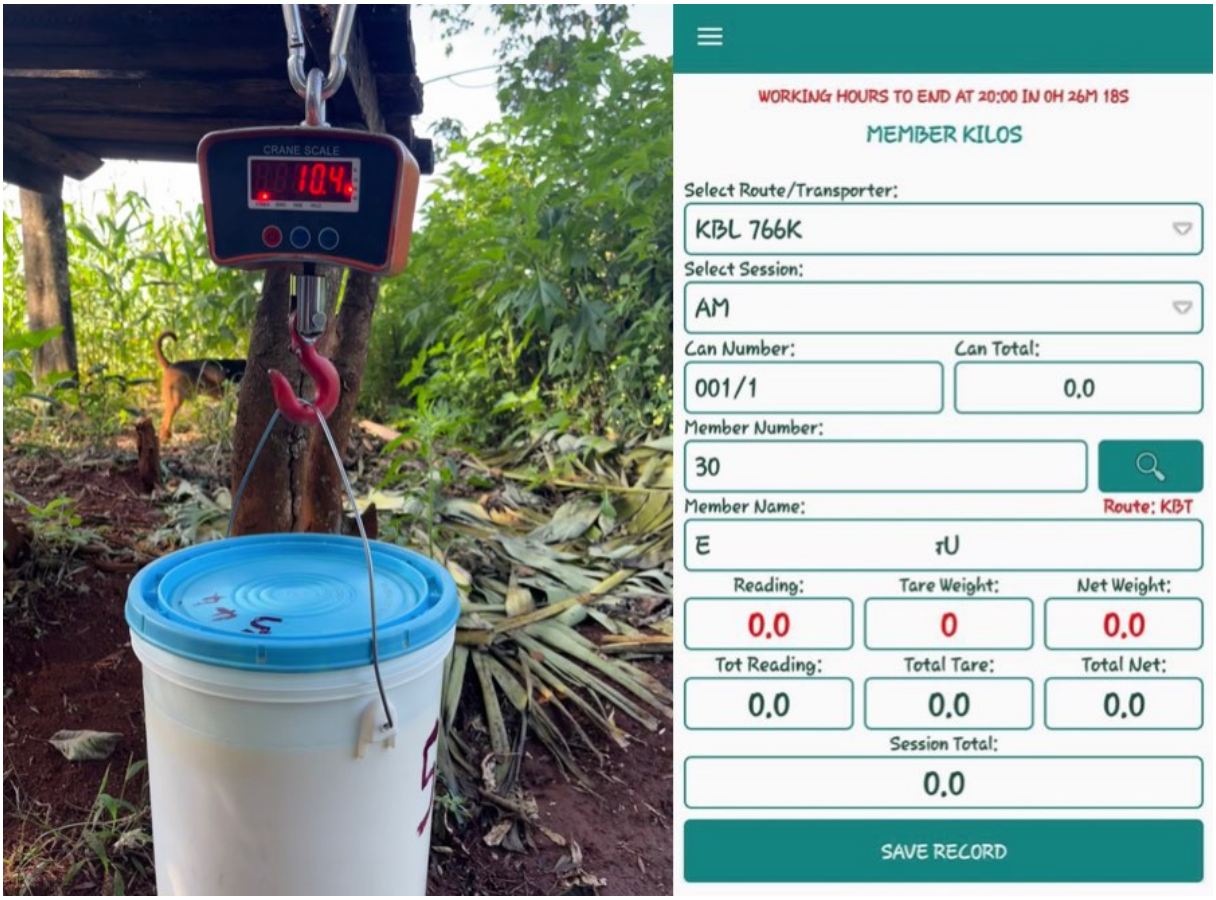


Notes: This figure illustrates the flow of milk through the formal value chain in Kenya's dairy sector, facilitated by cooperatives. Dairy cooperatives hire milk transporters to collect milk from farmers. Milk is not tested at this point due to the high testing cost (both time and monetary). Transporters usually aggregate the milk from multiple farmers to fill larger milk cans, which are then transported to collection centers. The collection centers pour the milk together into cooling plants (usually containing 10 to 100 of these cans) for sale to processors. These processors then perform comprehensive milk quality tests. Processors always classify milk into three categories based on the milk quality testing results: (1) accept the aggregated milk with a premium price, (2) accept it without a premium price, or (3) reject it. The reason is that high-quality milk can be used to produce high-value products like buttermilk and yogurt. Medium-quality milk is limited to use in ultra-pasteurized liquid packets, but low-quality milk cannot be traded by law. I rely on the transporters to collect traceability data using the digital app, as is shown in Figure A3.

Figure A2: Assigning Unique IDs to Each Aggregated Milk Container

Notes: This figure displays the aggregated cans prior to marking (top left panel) and after marking (bottom left panel).

Figure A3: Digital Traceability System



Notes: This figure demonstrates how transporters use the digital app to collect traceability data. The app records key information such as the collection route, whether the milk is from the morning or evening, aggregated can numbers, and corresponding trips (e.g., 001/1 indicates can 1 on the first trip). To prevent errors, the app displays the total can volume, ensuring it does not exceed 50 liters. Additionally, the app collects unique member IDs assigned by the cooperatives; when a member ID is selected, the member's name appears for verification. Transporters are also required to input the total weight and tare weight to calculate the net weight.

Figure A4: Pouring Milk into Aggregated Containers on Truck

Notes: When collecting milk, transporters pour it into aggregated containers on the truck and are advised to randomly distribute farmers' milk into different containers. This approach increases variation in can-level composition and enhances the model's performance. Before departing for milk collection, the aggregated containers are arranged sequentially according to their can IDs. In the field, milk from the first farmer is poured into the first can, the second farmer's milk into the second can, and so on. Once all the aggregated cans have been filled with milk from at least one farmer, a new round of pouring begins. During this round, transporters randomly select a starting can and then proceed sequentially through the containers. This process ensures that the neighboring farmers do not consistently mix their milk (e.g., the 1st and 2nd farmers), and it also prevents cases where the 1st and 23rd (if there are 22 cans in total) farmers' milk is always mixed in the same container.

Figure A5: Mixing Milk before Taking Samples at the Cooling Plant

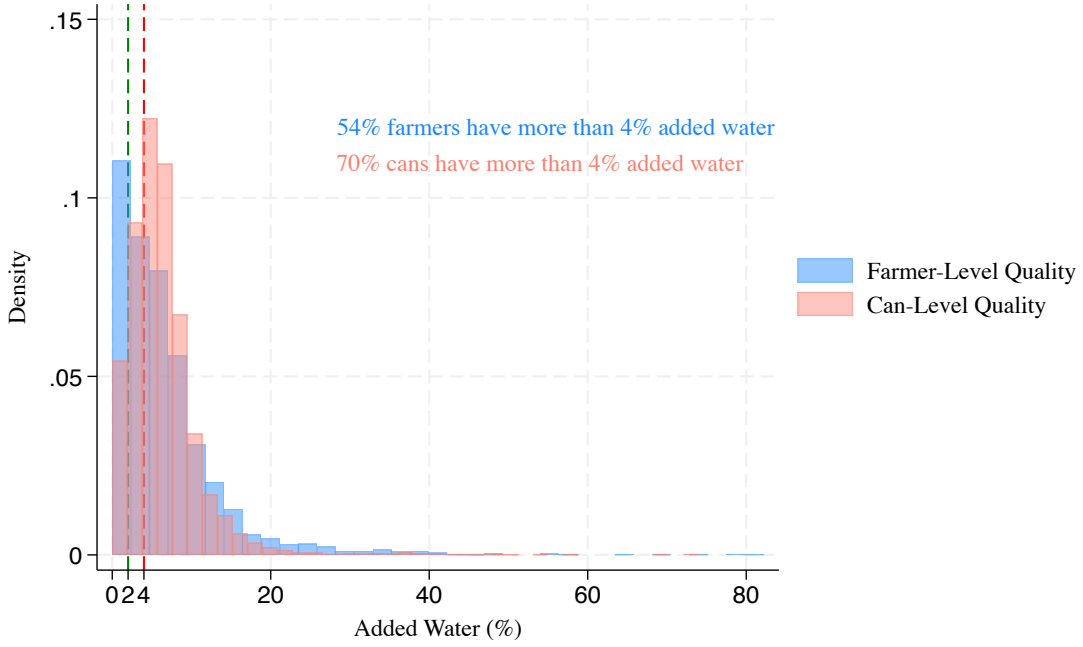
Notes: Before collecting samples from the aggregated cans, the field team used a plunger to thoroughly mix the milk by stirring it from top to bottom until it was well combined. This proved to be necessary and useful during the pilot trial, and the field team included this as one of the important steps in the milk sampling protocol.

Figure A6: Milk Samples with Ice Bricks

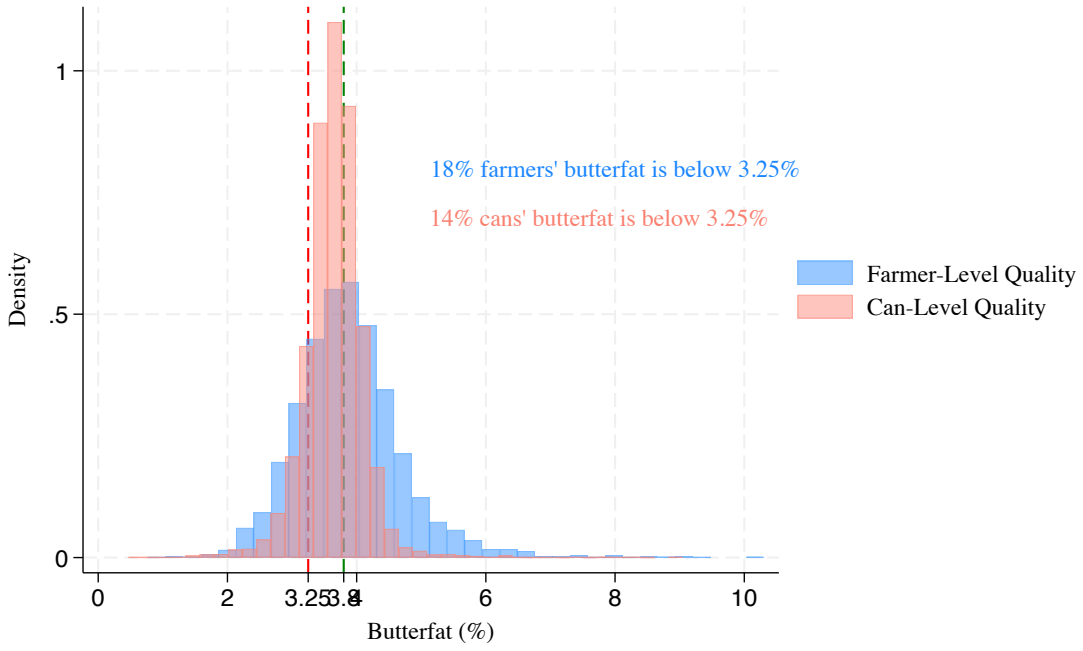
Notes: Milk sample is poured into small sample tubes with unique sample IDs and stored with ice bricks to keep fresh while shipping to the milk testing lab.

Figure A7: Milk Quality Distribution (Pooled All Rounds)

(a) Panel A: Added Water



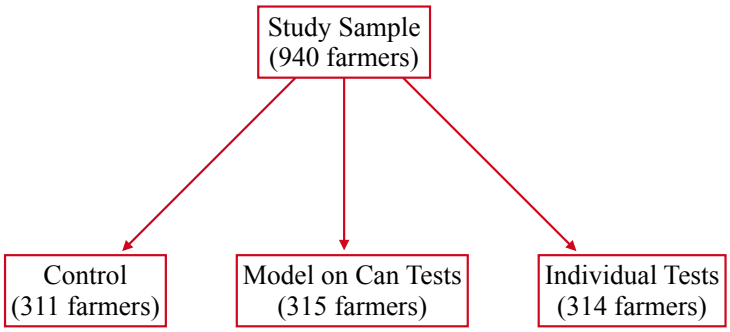
(b) Panel B: Butterfat



Notes: This figure shows the distributions of added water and butterfat at both the farmer and can levels. The added water is non-negative, and its distribution at the farmer level closely resembles a Half-Normal distribution, which I have selected as the prior for our analysis. Butterfat, being strictly positive, aligns well with a Gamma distribution at the farmer level, which I have chosen as the prior for this parameter. The national KeBS standards require that added water must be less than 4% and butterfat must exceed 3.25%.

Figure A8: Experimental Design and Timeline

(a) Panel A: Experimental Design



(b) Panel B: Timeline

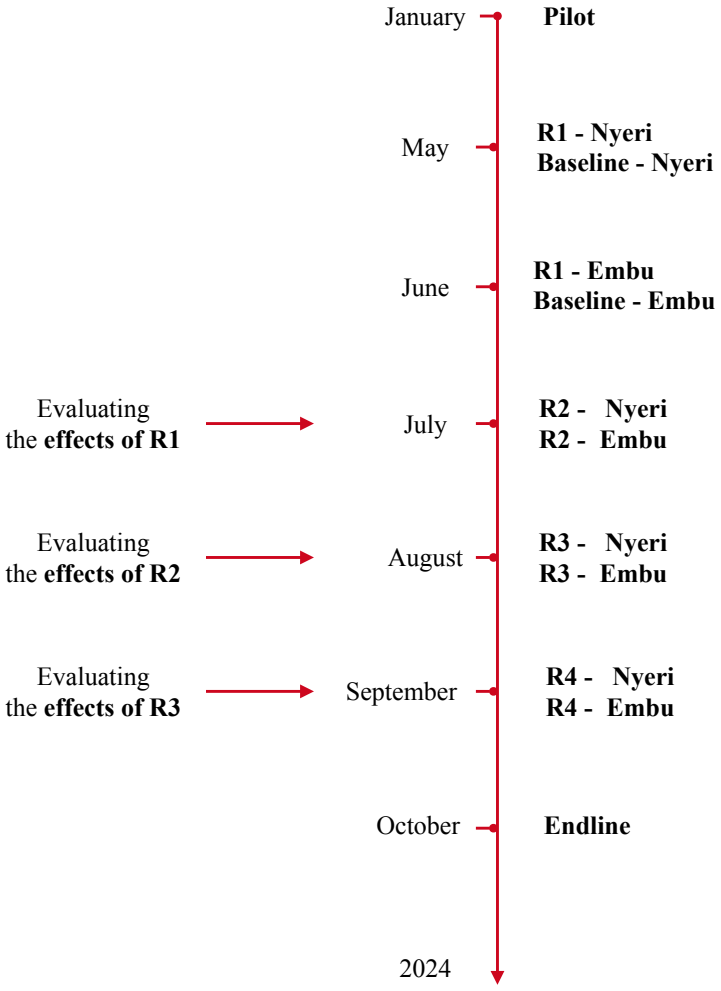


Figure A9: Added Water and Butterfat Tests

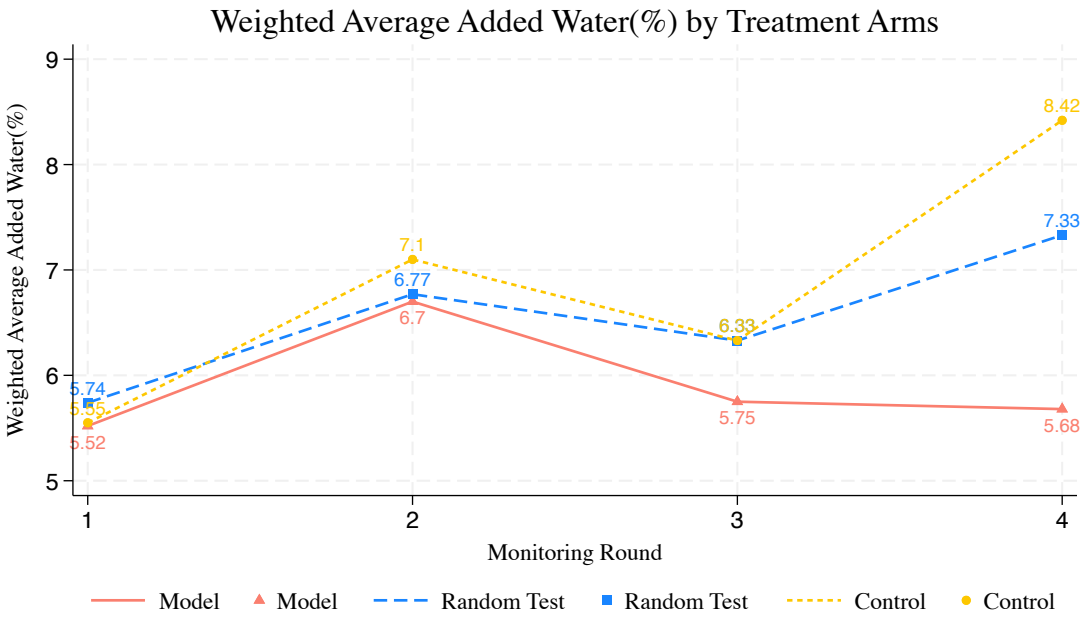
Notes: Added water and butterfat are tested by the *Ekomilk ULTRA PRO* ultrasonic milk analyzer, produced by *Eon Trading INC* in Bulgaria. It typically takes 1 minute per test. The lab staff usually tests 3 times per sample and records the last testing results to minimize the impact of the residual milk from the previous samples.

Figure A10: Bacteria Test

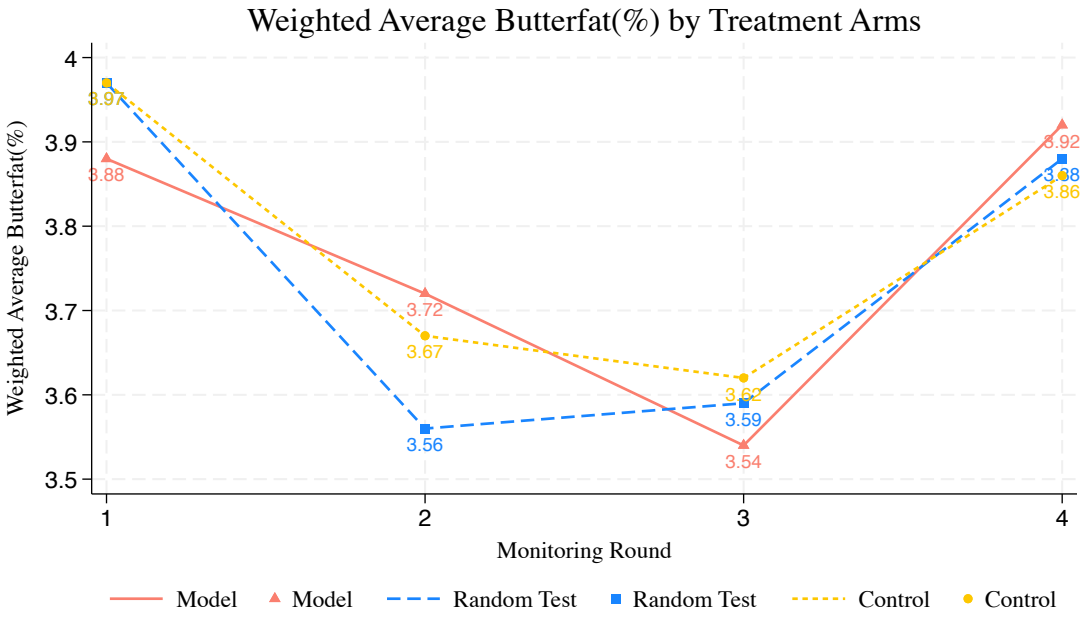
Notes: The bacteria is tested using the Resazurin Test (RT), in which a prepared solution of resazurin dye is added to the milk sample. The color change after incubation indicates the milk quality: blue for good quality, purple or pink for normal quality, and colorless for bad quality. In the RT test, 1 mL of 0.005% resazurin dye solution is added to 10 mL of the milk sample in a test tube and then incubated in a water bath at 37°C for 10 minutes. The test results are interpreted based on how quickly the color shifts, reflecting bacterial activity. Bacterial levels are determined by comparing the color to a reference chart.

Figure A11: Weighted Average Milk Quality over Time by Treatment Arms

(a) Added Water



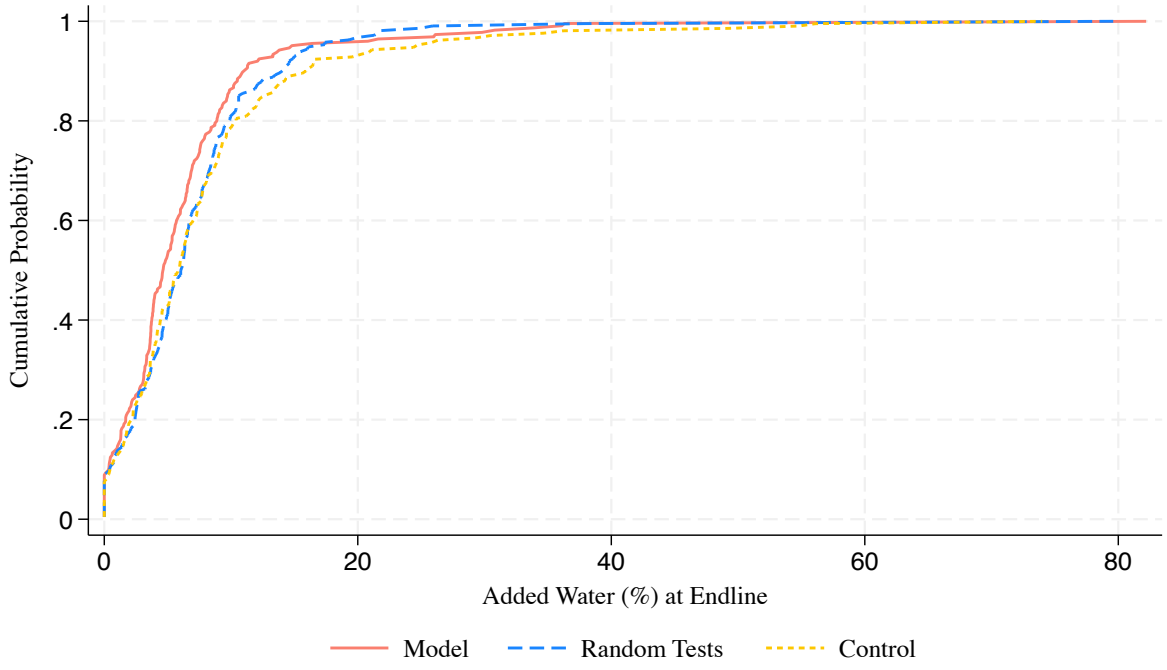
(b) Butterfat



Notes: This figure shows the weighted average milk quality (weighted by individual farmers' milk volumes) across monitoring rounds for different treatment arms. The upper panel illustrates added water, while the bottom panel depicts butterfat. These metrics are used to assess overall improvement within each treatment group and to conduct a cost-efficiency analysis.

Figure A12: Cumulative Distribution Function(CDF) of Endline Quality by Treatment Arms

(a) Added Water



(b) Butterfat

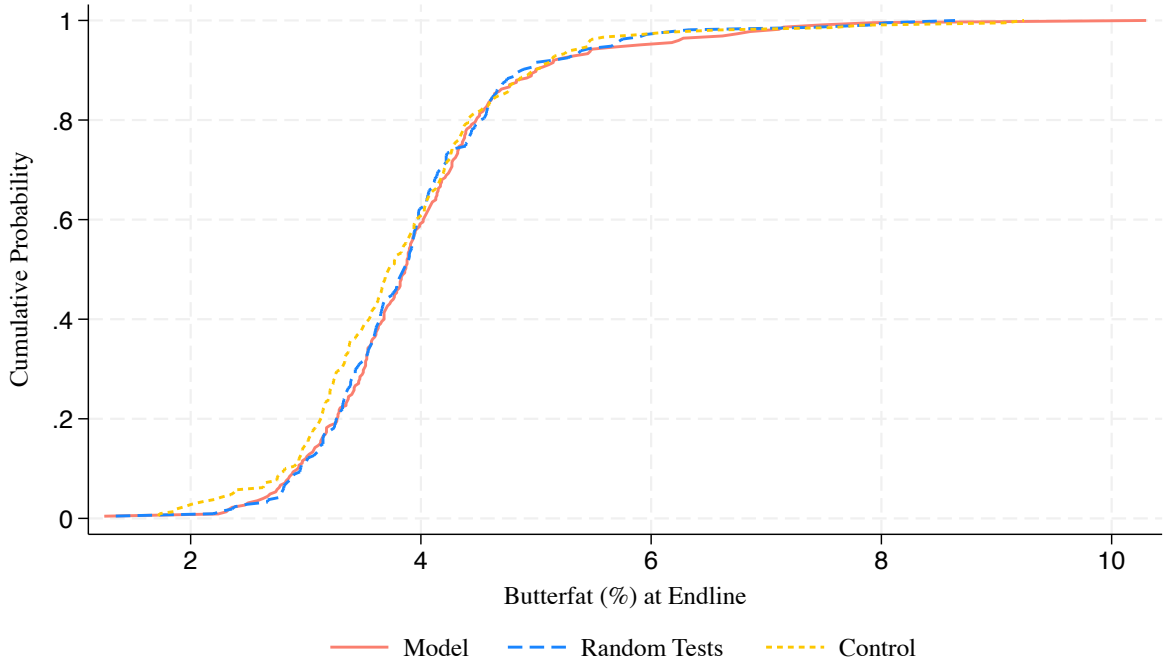
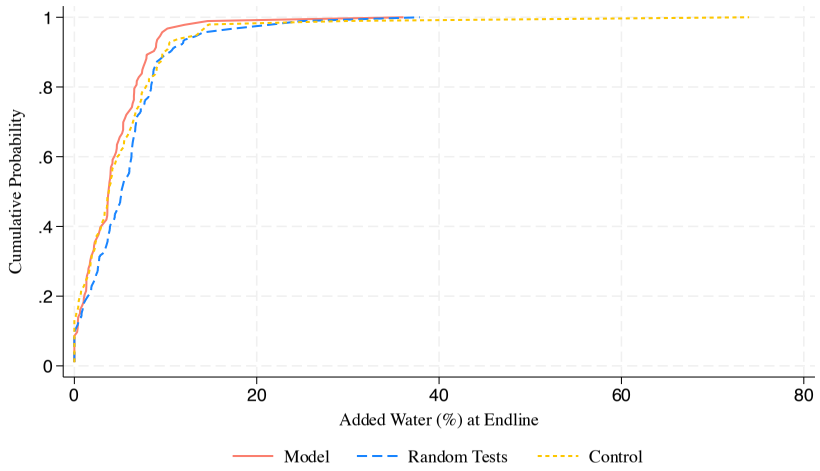
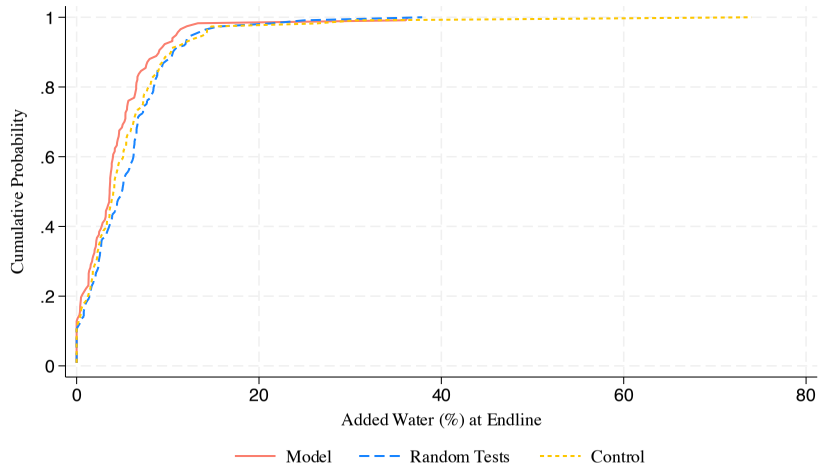


Figure A13: CDF of Endline Added Water (Subgroup Analysis)

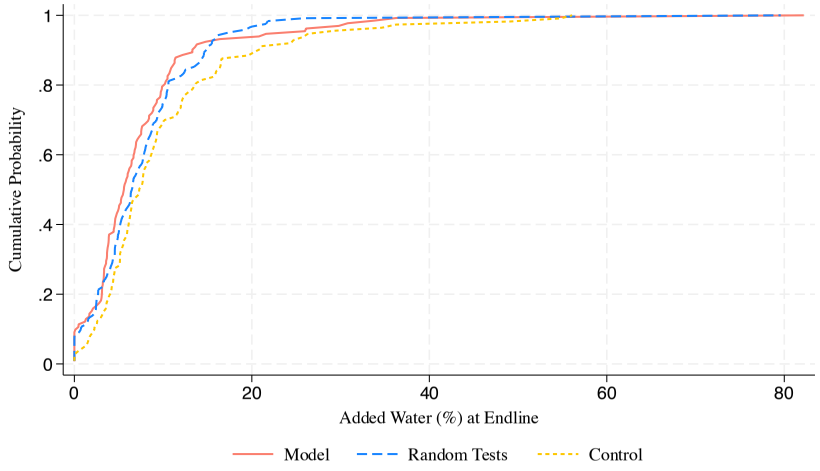
(a) Good Added Water by Model



(b) Good Added Water by Individual Tests



(c) Bad Added Water by Model



(d) Bad Added Water by Individual Tests

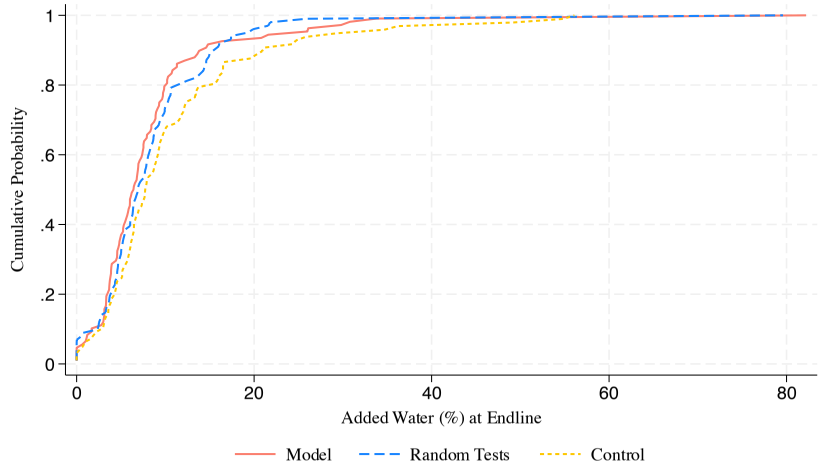
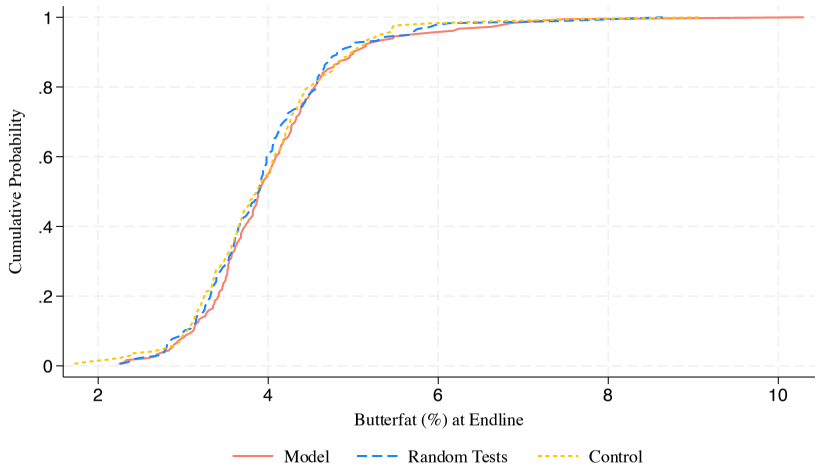
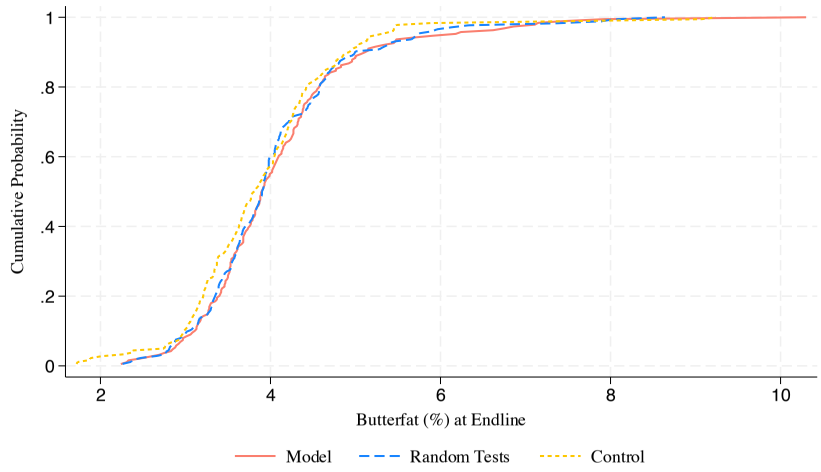


Figure A14: CDF of Endline Butterfat (Subgroup Analysis)

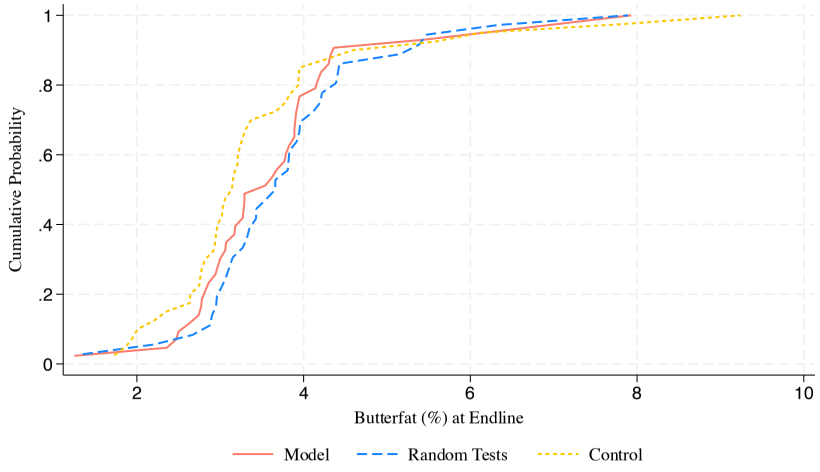
(a) Good Butterfat by Model



(b) Good Butterfat by Individual Tests



(c) Bad Butterfat by Model



(d) Bad Butterfat by Individual Tests

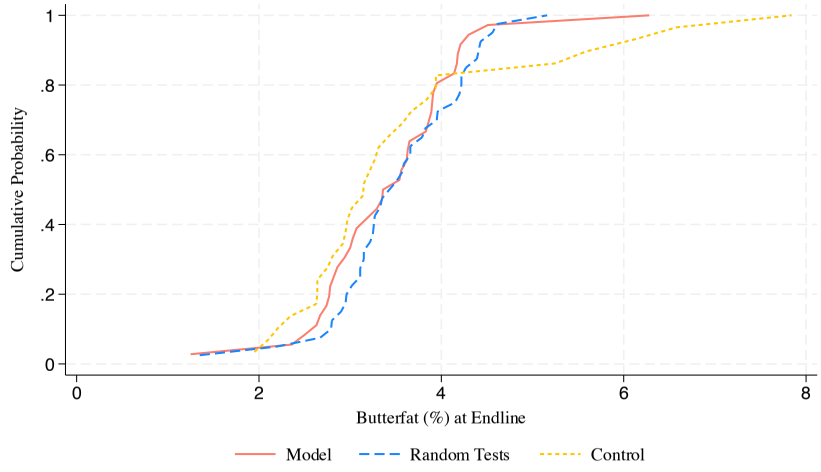


Table A1: Attrition Balance

	=1 if Completed Endline Survey	=1 if Completed Endline Milk Tests
	(1)	(2)
Panel A: Pooled Regression		
Random Tests (β)	0.02	0.01
	(0.04)	(0.04)
Model Detection (γ)	-0.01	0.04
	(0.04)	(0.04)
p-value:		
$\beta = \gamma$	0.477	0.371
Observations	719	940
Control Mean of Depvar	0.80	0.68
Control SD of Depvar	0.40	0.47
Panel B: Interaction with Baseline Water by Individual Tests		
Random Tests × Bad Water (β)	-0.06	0.02
	(0.07)	(0.07)
Random Tests	0.04	-0.00
	(0.05)	(0.05)
Model Detection × Bad Water (γ)	0.11	0.01
	(0.07)	(0.07)
Model Detection	-0.06	0.03
	(0.05)	(0.05)
=1 if Bad Water by Individual Tests	0.02	-0.00
	(0.05)	(0.05)
p-value:		
$\beta = \gamma$	0.024	0.938
Observations	719	940
Control × Good Baseline Water : Mean	0.79	0.68
Control × Good Baseline Water : SD	0.41	0.47
Panel C: Interaction with Baseline Water by Model		
Random Tests × Bad Water (β)	0.07	0.02
	(0.07)	(0.07)
Random Tests	-0.02	-0.00
	(0.06)	(0.06)
Model Detection × Bad Water (γ)	0.05	0.10
	(0.07)	(0.07)
Model Detection	-0.04	-0.02
	(0.06)	(0.06)
=1 if Bad Water by Model	-0.02	-0.08
	(0.05)	(0.05)
p-value:		
$\beta = \gamma$	0.760	0.309
Observations	719	940
Control × Good Baseline Water : Mean	0.80	0.73
Control × Good Baseline Water : SD	0.40	0.45

Notes: This table reports survey attrition and milk test attrition, along with balance tests across treatment arms and by baseline quality levels. Panel A presents results from the estimating equation (2), while Panels B and C report results from the estimating equation (3). For the regressors, "Bad Water" classifications are derived from individual-level random tests in Panel B and from model detection in Panel C. "Bad Water" indicates the failure to meet the KeBS standards for Added Water. All regressions include collection route-fixed effects. ***, **, and * represent significance at 1%, 5%, and 10% respectively.

Figure A15: Farmers' Milk Quality Training

Notes: Cooperatives contacted the regional *Brookside* trainers to provide training for farmers with low-quality milk, focusing on improving butterfat and reducing bacteria levels. Attendance sheets were signed by each participating farmer.

Table A2: Treatment Effects on Investment in Butterfat

	Concentrates Cost(USD) in the Past 30 Days	Dry Matters Percentage among the Total Feed
	(1)	(2)
Panel A: Pooled Regression		
Random Tests (β)	2.79	-2.59
	(3.09)	(1.68)
Model Detection (γ)	2.70	-2.27
	(3.12)	(1.70)
p-value:		
$\beta = \gamma$	0.978	0.850
Observations	574	574
Control Mean of Depvar	24.44	27.98
Control SD of Depvar	30.21	17.02
Panel B: Interaction with Baseline Butterfat by Individual Tests		
Random Tests × Bad Butterfat (β)	-6.97	2.09
	(8.63)	(4.69)
Random Tests	3.71	-2.75
	(3.38)	(1.84)
Model Detection × Bad Butterfat (γ)	-10.54	4.20
	(8.98)	(4.88)
Model Detection	4.22	-2.84
	(3.40)	(1.85)
=1 if Bad Butterfat by Individual Tests	6.98	-3.92
	(6.74)	(3.67)
p-value:		
$\beta = \gamma$	0.660	0.631
Observations	574	574
Control × Good Baseline Butterfat : Mean	24.44	28.47
Control × Good Baseline Butterfat : SD	30.21	17.56
Panel C: Interaction with Baseline Butterfat by Model		
Random Tests × Bad Butterfat (β)	-1.27	-6.12
	(7.89)	(4.28)
Random Tests	2.98	-1.45
	(3.45)	(1.87)
Model Detection × Bad Butterfat (γ)	1.87	0.58
	(8.08)	(4.38)
Model Detection	2.30	-2.36
	(3.48)	(1.89)
=1 if Bad Butterfat by Model	-2.23	0.38
	(5.60)	(3.04)
p-value:		
$\beta = \gamma$	0.701	0.131
Observations	574	574
Control × Good Baseline Butterfat : Mean	24.44	27.91
Control × Good Baseline Butterfat : SD	30.21	17.89

Notes: This table presents results from running equation (2) for Panel A, and equation (3) for Panel B and Panel C. Regressions include baseline measurements of outcome, collection route fixed effects, and baseline controls for the number of household members and the number of years keeping cattle. For the regressors, "Bad Butterfat" classifications are derived from individual-level random tests in Panel B and from model detection in Panel C. "Bad Butterfat" indicates the failure to meet the KeBS standards for Butterfat. ***, **, and * represent significance at 1%, 5%, and 10% respectively.

Table A3: Treatment Effects on Investment in Bacteria

	=1 if Use Aluminium Milk Cans	=1 if Mix Morning and Evening Milk	=1 if Use Milking Jelly after Milking	=1 if Use Different Towels for Milking Different Cows
	(1)	(2)	(3)	(4)
Panel A: Pooled Regression				
Random Tests (β)	0.02	-0.00	0.12**	0.08*
	(0.04)	(0.05)	(0.05)	(0.04)
Model Detection (γ)	-0.01	0.04	0.08	0.04
	(0.04)	(0.05)	(0.05)	(0.04)
p-value:				
$\beta = \gamma$	0.498	0.391	0.466	0.385
Observations	574	574	574	470
Control Mean of Depvar	0.74	0.41	0.57	0.73
Control SD of Depvar	0.44	0.49	0.50	0.45
Panel B: Interaction with Baseline Quality by Individual Tests				
Random Tests × Bad Bacteria (β)	0.01	-0.00	0.00	0.11
	(0.11)	(0.12)	(0.12)	(0.11)
Random Tests	0.02	-0.01	0.11**	0.05
	(0.05)	(0.06)	(0.06)	(0.05)
Model Detection × Bad Bacteria (γ)	-0.04	0.01	0.10	-0.02
	(0.11)	(0.13)	(0.13)	(0.11)
Model Detection	0.00	0.04	0.06	0.05
	(0.05)	(0.06)	(0.05)	(0.05)
=1 if Bad Bacteria by Individual Tests	-0.05	0.05	0.04	0.00
	(0.08)	(0.09)	(0.09)	(0.08)
p-value:				
$\beta = \gamma$	0.595	0.932	0.424	0.242
Observations	574	574	574	470
Control × Good Baseline Bacteria : Mean	0.74	0.39	0.57	0.76
Control × Good Baseline Bacteria : SD	0.44	0.49	0.50	0.43

Notes: This table presents results from running equation (2) for Panel A, and equation (3) for Panel B. Regressions include baseline measurements of outcome, collection route fixed effects, and baseline controls for the number of household members and the number of years keeping cattle. For the regressors, "Bad Bacteria" classifications are derived from individual-level random tests in Panel B. "Bad Bacteria" indicates the failure to meet the KeBS standards for Bacteria. ***, **, and * represent significance at 1%, 5%, and 10% respectively.

B Discussion of Welfare

The selection criteria for the partner cooperatives in this study include the willingness to set up the traceability systems and, most importantly, the agreement that they would not take actions that might hurt farmers once I share the quality information. Both cooperatives I recruited to the study have promised these at the beginning of the project, although the research team cannot fully control how the cooperatives react to the quality information. Therefore, I recognized in the IRB[27] application that, in the worst case, there is a risk for farmers to be rejected by cooperatives because of the low milk quality.[28] I have also fully notified farmers of these potential risks, and farmers' participation in the study is voluntary. I asked for consent from farmers for the milk testing[29] and surveys separately. I have also informed farmers that they could choose to drop out anytime during the study. As a result, 30 percent of farmers dropped out of the testing halfway.

At the same time, the research team closely monitored the cooperatives' responses after each round of information sharing. I observed that cooperatives started training low-quality (butterfat and bacteria) farmers and talked to farmers about water adulteration since I started the information sharing. Both cooperatives chose to have a conversation with low-quality farmers first without any real consequences during the first two rounds of quality monitoring and information sharing. The cooperatives withheld a portion of the milk payment from a small group of farmers who repeatedly diluted their milk with water after the third round of information sharing. Once I observed this, I stopped the information sharing immediately[30]. After the conversation with the cooperatives, both agreed to fully restore all payments to the farmers. The ultimate goal of cooperatives is to let members take milk quality seriously, not necessarily to really punish farmers.

Note that any information-revealing study[31] aiming to reduce market inefficiencies will unavoidably lead to the reallocation of the welfare of involved parties, and the more informed

[27]The experiment and data collection were approved by the UCSC IRB, the International Livestock Research Institute (ILRI) Institutional Research Ethics Committee (IREC), as well as the in-country research permit from the National Commission for Science, Technology, and Innovation (NACOSTI).

[28]In reality, the cooperatives have never rejected farmers' milk because of information sharing in this study since the information sharing happened after they had already purchased the milk from farmers.

[29]Originally, I was concerned that the consent for milk testing would create selection issues and have unbalanced treatment and control groups, but a trial with ten good-type farmers and ten bad farmers in the pilot showed that the consent rate was almost 90%, and it's balanced between good- and bad-type farmers, then I included the information in the IRB application and rolled out to the full project. The attrition balance for the formal project can be found in Table A1.

[30]The original plan was to share four rounds of information in total, then there would be a final quality check, but due to the observed incidence, I did not share the fourth round quality monitoring results, and I used the fourth round results as the final outcomes of the project.

[31]Including Kelley et al. (2024) about installing GPS trackers on Kenyan matatu to inform the matatu owners of the hired drivers' driving behavior, as well as the companion paper Lane et al. (2024), which informs the passengers about drivers' driving behavior.

agent will have less information rents as a result of information revealing in the short run. However, this does not have to be the case, nor does it suggest they will be worse off in the long run. Farmers can either increase their income by diluting milk with water or achieve the same result by stopping the practice, which benefits everyone with higher prices. By improving milk quality, the cooperatives can earn a 6% price bonus, which can be passed on to the farmers. Choosing to stop adding water is a simple change that can happen quickly. In the long run, farmers whose milk is of insufficient quality might not be aware of their milk quality and simply do not know how to produce high-quality milk. It is a good chance to get follow-up guidance from the cooperatives. It is reasonable to believe that cooperatives will give them guidance and even pay them bonuses if the milk quality is improved. In fact, I have already observed that our partner cooperatives increased the credit limits for those farmers who have never added water to the milk and need money to buy animal feeds, as the cooperative believes that these farmers are more trustworthy and will be more likely to pay back the loans. The cooperative is also willing to see an increase in the quantity of "good" type of milk. In summary, farmers could benefit from the follow-up guidance, potential quality bonuses, and improved access to credit via built trust in the long run.

It is also important to note that if the milk quality is "bad," both farmers and cooperatives should know it, as bad milk with quality below standards set by the KeBS is not allowed to be traded by law, and it will bring a risk to public health. Even for added water alone, there is a significant positive correlation between adding water and bacteria, as is shown in Table A4, as the water added to the milk would not necessarily be clean, purified water. In Table A4, Panel A indicates that a 1% increase in added water percentage is associated with a 1% higher likelihood of the milk being classified as "Bad Bacteria" according to national KeBS standards, a relationship that is statistically significant at the 1% level. Panel B shows that being classified as "Bad Quality" in terms of added water is associated with a 2% higher likelihood of the milk being classified as "Bad Bacteria".

Farmers sell the milk not only to cooperatives but also to their neighbors and local schools. There is a public health concern for this unpasteurized milk to be traded and consumed in the local area. Milk consumers will benefit from the improved milk market in the long run. At the endline of the project, 99% of farmers agreed that milk should always be tested by buyers, and 90% of farmers agreed that milk should be paid based on quality, as is shown in Table A5.

Table A4: Correlation between Added Water and Bacteria

	=1 if Bad Bacteria
	(1)
Panel A: Absolute Level of Added Water	
Added Water (%)	0.01***
	(0.00)
Observations	3574
Mean of Depvar	0.14
SD of Depvar	0.35
Panel B: Bad Added Water Classification	
=1 if Bad Added Water	0.02*
	(0.01)
Observations	3574
Control Mean of Depvar	0.13
Control SD of Depvar	0.34

Notes: Pooled all four rounds of individual milk samples. Panel A indicates that a 1% increase in added water percentage is associated with a 1% higher likelihood of the milk being classified as "Bad Bacteria" according to national KeBS standards, a relationship that is statistically significant at the 1% level. Panel B shows that being classified as "Bad Quality" in terms of added water is associated with a 2% higher likelihood of the milk being classified as "Bad Bacteria". ***, **, and * represent significance at 1%, 5%, and 10% respectively.

Table A5: Farmers' Perception Regarding Milk Tests and Quality Based Payment

	Mean	SD	N
Milk Should Always be Tested by Buyers	0.99	0.10	574
Milk Price Should be Based on Quality (Baseline)	0.61	0.49	719
Milk Price Should be Based on Quality (Endline)	0.90	0.29	574

C Technical Appendix

As described in the main text, I use Bayesian statistical models to predict milk quality. This appendix provides the details for estimation. According to Bayes' Theorem, the posterior distribution for the milk quality parameter ρ follows:

$$f(\rho \mid \text{data}) = \frac{f(\text{data} \mid \rho) f(\rho)}{f(\text{data})},$$

where $f(\text{data} \mid \rho)$ is the Likelihood function, $f(\rho)$ is the prior distribution for the parameter, and $f(\text{data})$ is the marginal distribution of the data. I will explain each component of the Bayesian model for different quality parameters one by one. The basic idea is that we incorporate prior information for milk quality across different farmers and use their contribution toward different aggregate cans and the quality measurements of these cans to infer individual quality averages ex-post.

C.1 Model Setup for Added Water

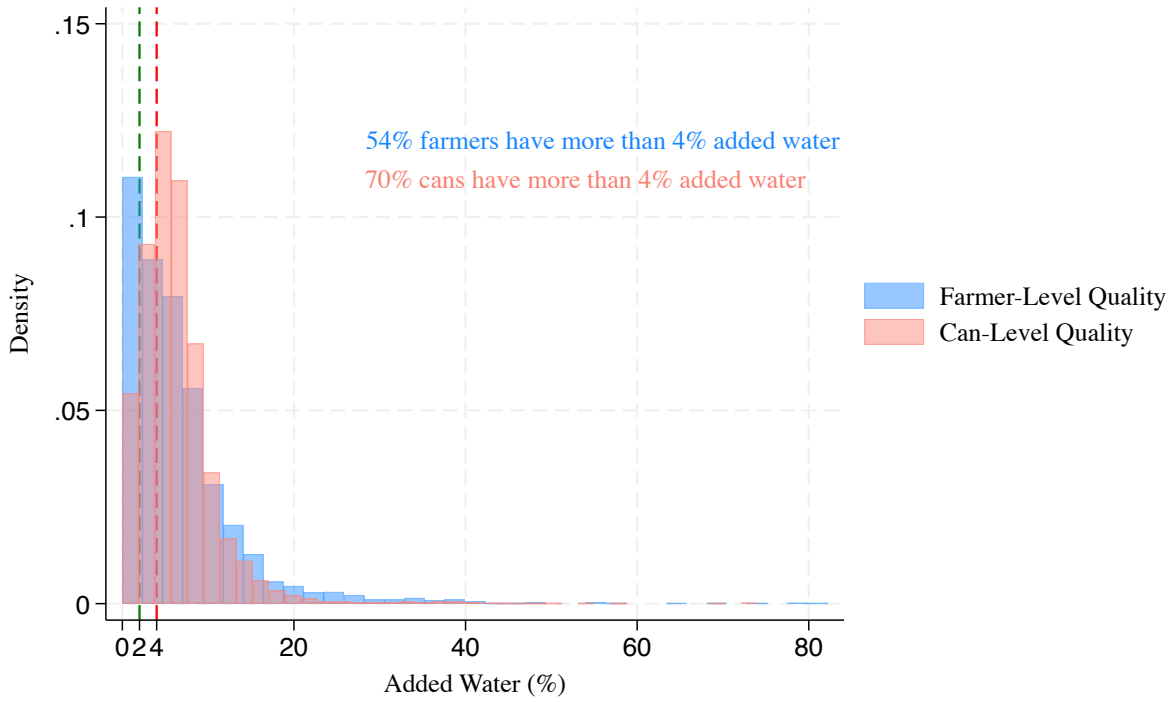
Let us assume there are n different farmers, and the added water in their individual milk is represented by the parameters $\rho_1^{indiv}, \rho_2^{indiv}, \ldots, \rho_n^{indiv}$.

C.1.1 Priors for Added Water

The individual added water level, ρ_i^{indiv}, is non-negative by definition. Based on the histogram of the observed data in Figure A16, I model the potential prior distribution of added water using the half-normal distribution:

$$\rho_i^{indiv} \sim \text{Normal}^+(0, (\sigma^{indiv})^2)$$

Figure A16: Added Water Distribution (Pooled All Rounds)



Notes: This figure shows the distributions of added water at both the farmer and can levels. The added water is non-negative, and its distribution at the farmer level closely resembles a Half-Normal distribution, which I have selected as the prior for our analysis. The national KeBS standards require that added water must be less than 4%.

Mathematically, the probability density function (PDF) of the positive half-normal distribution is given by:

$$f(\rho_i^{indiv}; 0, \sigma^{indiv}) = \begin{cases} \frac{\sqrt{2}}{\sigma^{indiv}\sqrt{\pi}} \exp\left(-\frac{(\rho_i^{indiv})^2}{2(\sigma^{indiv})^2}\right), & \rho_i^{indiv} \geqslant 0 \\ 0, & \text{otherwise.} \end{cases}$$

Without additional information, the prior of σ^{indiv} is chosen to be an inverse gamma distribution:

$$\sigma^{indiv} \sim \text{InvGamma}(3, 1)$$

This is an innocuous assumption that simply assigns a random starting value for the dispersion in the Half-Normal PDF.

C.1.2 Likelihood for Added Water

Theoretical Can Level Quality

Given the share of the individual farmer i's milk in the can j, w_{ij}^{data}, the aggregated added water level, ρ_j^{can}, is a weighted average of the individual added water from m_j different farmers' milk in can j:

$$\rho_j^{can} = \sum_{i=1}^{m_j} w_{ij}^{data} \rho_i^{indiv}$$

Thus, for all the aggregated cans, we have:

$$\boldsymbol{\rho}_{\boldsymbol{aw,J\times 1}}^{\boldsymbol{can}} = W_{\mathbf{J\times n}} \boldsymbol{\rho}_{\boldsymbol{aw,n\times 1}}^{\boldsymbol{indiv}},$$

where $W_{\mathbf{J\times n}} = \begin{bmatrix} w_{11} & w_{12} & \dots & w_{1n} \\ w_{21} & w_{22} & \dots & w_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ w_{(J)1} & w_{(J)2} & \dots & w_{(J)n} \end{bmatrix}$. Here, **W** is the weight matrix with **J rows** since we have J cans, and **n columns** as we have n different farmers. $\boldsymbol{\rho}_{\boldsymbol{aw,n\times 1}}^{\boldsymbol{indiv}} = \begin{bmatrix} \rho_{aw,1}^{indiv} \\ \rho_{aw,2}^{indiv} \\ \vdots \\ \rho_{aw,n}^{indiv} \end{bmatrix}$, which is the individual added water(aw) matrix with n rows. $\boldsymbol{\rho}_{\boldsymbol{aw,J\times 1}}^{\boldsymbol{can}} = \begin{bmatrix} \rho_{aw,1}^{can} \\ \rho_{aw,2}^{can} \\ \vdots \\ \rho_{aw,J}^{can} \end{bmatrix}$, which is the can level added water matrix with J rows.

Data Model

The aggregated can-level added water, y_j^{data}, is typically observed with measurement errors around its theoretical mean ρ_j^{can}. Further, since y_j^{data} is also non-negative, I model the distribution of y_j^{data} as a normal distribution truncated below from zero, with location parameter the theoretical mean ρ_j^{can} and scale parameter the can-level measurement error σ^{can}:

$$y_j^{data} \sim \text{Normal}^{+}(\rho_j^{can}, (\sigma^{can})^2),$$

where $\rho_j^{can} = \text{E}[y_j^{data}] = \sum_{i=1}^{m_j} w_{ij}^{data} \rho_i^{indiv}$. σ^{can} is assigned an inverse gamma prior:

$$\sigma^{can} \sim \text{InvGamma}(3, 1)$$

Given the observed y_j^{data}, known weights w_{ij}^{data}, the likelihood function below informs us how probable the observed aggregated added water is for different combinations of individual densities $\rho_1^{indiv}, \rho_2^{indiv}, \ldots, \rho_n^{indiv}$:

$$L\left(\mathbf{Y}^{data} \mid \rho_1^{indiv}, \rho_2^{indiv}, \ldots, \rho_n^{indiv}\right) = \prod_{j=1}^{J} L\left(y_j^{data} \mid \rho_1^{indiv}, \rho_2^{indiv}, \ldots, \rho_n^{indiv}\right)$$

$$= \prod_{j=1}^{J} \frac{\frac{1}{\sigma^{can}\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y_j^{data} - \rho_j^{can}}{\sigma^{can}}\right)^2}}{1 - \Phi\left(\frac{-\rho_j^{can}}{\sigma^{can}}\right)} = \prod_{j=1}^{J} \frac{\frac{1}{\sigma^{can}\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y_j^{data} - \sum_{i=1}^{m_j} w_{ij}^{data} \rho_i^{indiv}}{\sigma^{can}}\right)^2}}{1 - \Phi\left(\frac{-\sum_{i=1}^{m_j} w_{ij}^{data} \rho_i^{indiv}}{\sigma^{can}}\right)}$$

C.1.3 Posteriors for Added Water

Given the above setup, Bayes' theorem gives:

$$P\left(\rho_1^{indiv}, \rho_2^{indiv}, \ldots, \rho_n^{indiv} \mid \mathbf{Y}^{data}\right) \propto$$

$$L\left(\mathbf{Y}^{data} \mid \rho_1^{indiv}, \rho_2^{indiv}, \ldots, \rho_n^{indiv}\right) \times P\left(\rho_1^{indiv}\right) \times P\left(\rho_2^{indiv}\right) \times \ldots \times P\left(\rho_n^{indiv}\right)$$

$$= \prod_{j=1}^{J} L\left(y_j^{data} \mid \rho_1^{indiv}, \rho_2^{indiv}, \ldots, \rho_n^{indiv}\right) \times \prod_{i=1}^{n} P\left(\rho_i^{indiv}\right)$$

$$= \prod_{j=1}^{J} \frac{\frac{1}{\sigma^{can}\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y_j^{data} - \rho_j^{can}}{\sigma^{can}}\right)^2}}{1 - \Phi\left(\frac{-\rho_j^{can}}{\sigma^{can}}\right)} \times \prod_{i=1}^{n} \frac{2}{\sqrt{2\pi(\sigma^{indiv})^2}} \exp\left(-\frac{(\rho_i^{indiv})^2}{2(\sigma^{indiv})^2}\right)$$

$$= \prod_{j=1}^{J} \frac{\frac{1}{\sigma^{can}\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y_j^{data} - \sum_{i=1}^{m_j} w_{ij}^{data} \rho_i^{indiv}}{\sigma^{can}}\right)^2}}{1 - \Phi\left(\frac{-\sum_{i=1}^{m_j} w_{ij}^{data} \rho_i^{indiv}}{\sigma^{can}}\right)} \times \prod_{i=1}^{n} \frac{2}{\sqrt{2\pi(\sigma^{indiv})^2}} \exp\left(-\frac{(\rho_i^{indiv})^2}{2(\sigma^{indiv})^2}\right)$$

For the posterior full conditional distribution of each ρ_i^{indiv}, we have the following:

$$P\left(\rho_i^{indiv} \mid \rho_{-i}^{indiv}, \mathbf{Y}^{data}\right) \propto P\left(\rho_1^{indiv}, \rho_2^{indiv}, \ldots, \rho_n^{indiv} \mid \mathbf{Y}^{data}\right)$$

$$\propto \prod_{j=1}^{J} \frac{\frac{1}{\sigma^{can}\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y_j^{data}-\sum_{i=1}^{m_j} w_{ij}^{data}\rho_i^{indiv}}{\sigma^{can}}\right)^2}}{1-\Phi\left(\frac{-\sum_{i=1}^{m_j} w_{ij}^{data}\rho_i^{indiv}}{\sigma^{can}}\right)} \times \prod_{i=1}^{n} \frac{2}{\sqrt{2\pi(\sigma^{indiv})^2}} \exp\left(-\frac{(\rho_i^{indiv})^2}{2(\sigma^{indiv})^2}\right)$$

$$\propto \prod_{j=p}^{k} \frac{e^{-\frac{1}{2}\left(\frac{y_j^{data}-\sum_{i=1}^{m_j} w_{ij}^{data}\rho_i^{indiv}}{\sigma^{can}}\right)^2}}{1-\Phi\left(\frac{-\sum_{i=1}^{m_j} w_{ij}^{data}\rho_i^{indiv}}{\sigma^{can}}\right)} \times \exp\left(-\frac{(\rho_i^{indiv})^2}{2(\sigma^{indiv})^2}\right)$$

$$\propto \frac{\exp\left\{-\frac{1}{2}\left[\sum_{j=p}^{k}\left(y_j^{data}-\sum_{i=1}^{m_j} w_{ij}^{data}\rho_i^{indiv}\right)^2+\left(\rho_i^{indiv}\right)^2\right)\right]\right\}}{\prod_{j=p}^{k}\left[1-\Phi\left(\frac{-\sum_{i=1}^{m_j} w_{ij}^{data}\rho_i^{indiv}}{\sigma^{can}}\right)\right]},$$

where p and k mean that individual farmer i show up in can p, can $p+1$, ..., can k. Note that any individual i showed up at least one time, and individual $i's$ milk was divided into (k+1) different aggregated containers on that single day. We also have $\Phi\left(\frac{-\sum_{i=1}^{m_j} w_{ij}^{data}\rho_i^{indiv}}{\sigma^{can}}\right) = \int_{-\infty}^{\left(\frac{-\sum_{i=1}^{m_j} w_{ij}^{data}\rho_i^{indiv}}{\sigma^{can}}\right)} \phi(u)du$, $\phi(u)$ is the pdf of standard normal distribution: $\phi(u) = \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}u^2}$.

C.2 Estimation and Model Performance for Added Water

C.2.1 Estimation - Hamiltonian Monte Carlo (HMC) for Added Water

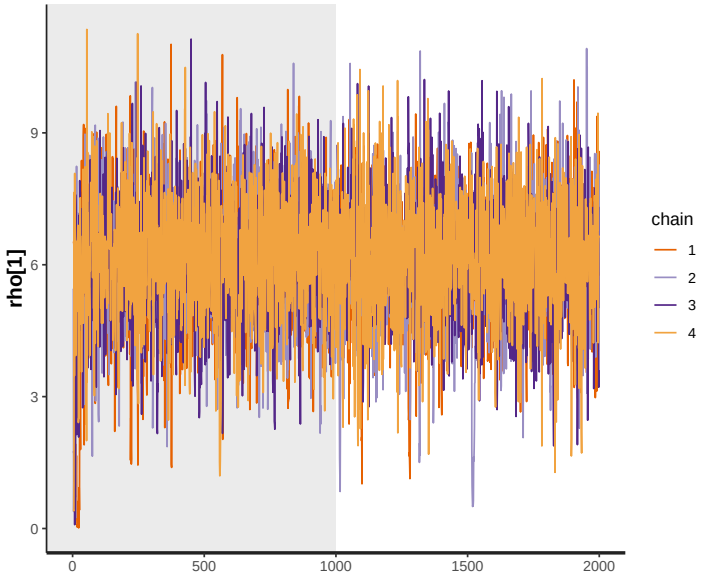
I implement the Bayesian model using RStan, which uses a modified version of the Hamiltonian Monte Carlo (HMC) algorithm[32] for Bayesian Inference.

I obtained point estimates and credible intervals for each ρ_i^{indiv}, σ^{indiv}, and σ^{can}. I attach the model estimation based on the modified Hamiltonian Monte Carlo (HMC) simulations[33] in Figure A17. I have also compared the Bayesian point estimates(posterior mean) to the Individual test results among 940 dairy farmers from cooperatives in two different counties in Figure A18.

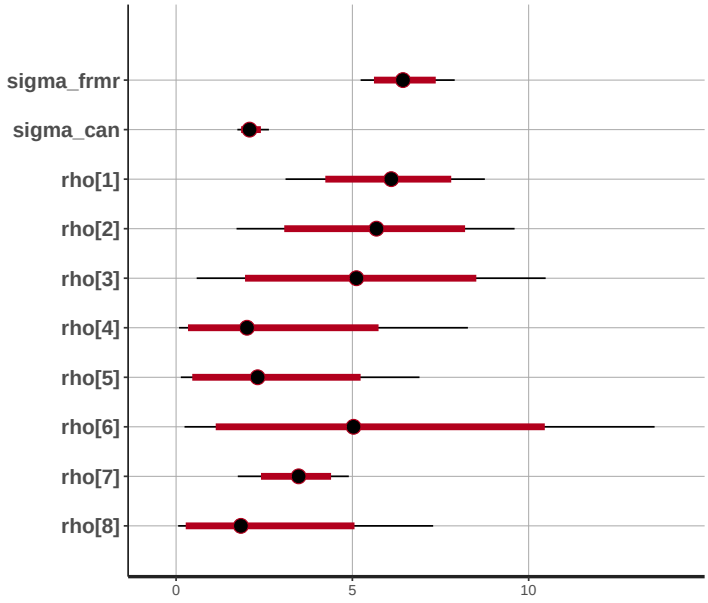
[32] Metropolis-Hastings within Gibbs Sampling will also work in this setting.
[33] Specifically, No-U-Turn sampler.

Figure A17: Hamiltonian Monte Carlo (HMC) Simulations for Added Water Estimation

(a) Panel A: Trace Plot

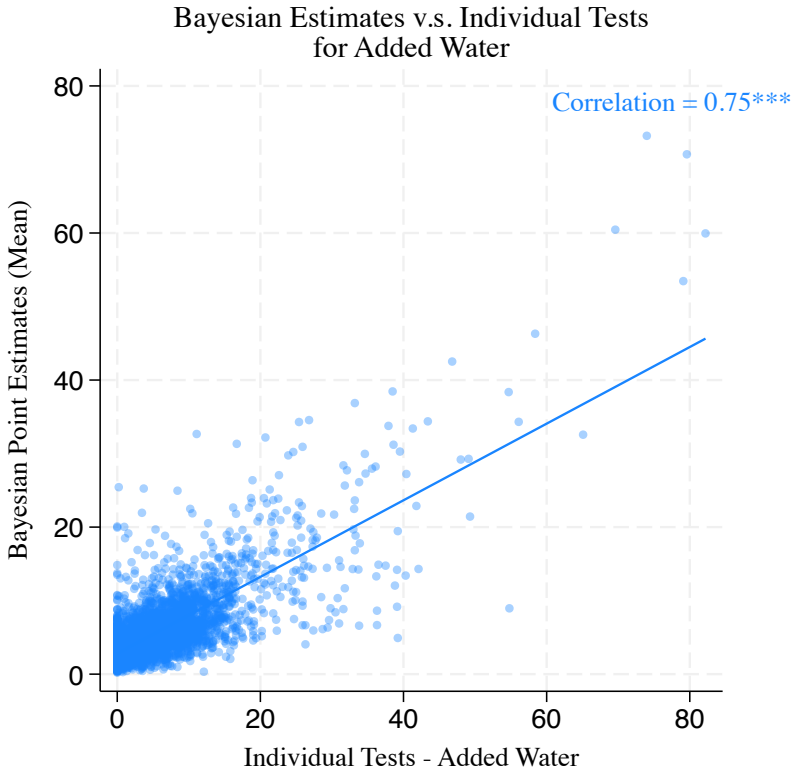


(b) Panel B: Caterpillar Plot



Notes: This figure presents the model estimation results based on the modified Hamiltonian Monte Carlo (HMC) simulations. Panel A displays the trace plots of the Markov chains, illustrating the time series of posterior draws. The first 1,000 draws, shaded in gray, represent the warmup phase, while the subsequent 1,000 draws from the post-warmup phase are used to approximate the posterior distribution of the parameter. Four independent chains were used, and their mixing was verified to ensure proper convergence during the post-warmup phase. Panel B presents the point estimates (black dots) alongside the posterior credible intervals, with the inner interval representing 80% credibility and the outer interval representing 95%, to illustrate the uncertainty in the estimates.

Figure A18: Comparison between Model Prediction and Individual Tests on Added Water



Notes: This figure compares Bayesian point estimates of average added water in the milk over the monitoring period with random individual test results. Both the X-axis and Y-axis represent absolute quality levels measured as percentages. The Pearson correlation coefficients between the Bayesian point estimates and the individual test results are calculated, pooling data across all four rounds of quality monitoring. ***, **, and * represent significance at 1%, 5%, and 10% respectively.

C.2.2 Posterior Probability Classification for Added Water

I then use the entire posterior distribution to classify farmers into three categories based on the minimally required standards posted by the government (Kenya Bureau Of Standards (KEBS)) and the processors' quality rewarding standards. To illustrate the idea, let $f\left(\rho_i^{indiv}\right)$ denote the posterior density function for farmer i. The probabilities for each tier[34] are calculated as follows:

- Probability of "Bad" Tier: $P_{Bad}\left(i\right) = \int_{\theta_{high}}^{\infty} f\left(\rho_i^{indiv}\right) d\rho_i^{indiv}$

- Probability of "Normal" Tier: $P_{Normal}\left(i\right) = \int_{\theta_{low}}^{\theta_{high}} f\left(\rho_i^{indiv}\right) d\rho_i^{indiv}$

[34]Note that for added water, the lower, the better.

- Probability of "Good" Tier: $P_{Good}(i) = \int_0^{\theta_{low}} f\left(\rho_i^{indiv}\right) d\rho_i^{indiv}$

where, $\theta_{low} = 2\%$, the processor's standards for quality rewarding, and $\theta_{high} = 4\%$, the minimal required standards by the KeBS.

I compare the predicted categories and categories based on a one-time test in the confusion matrix below and calculate the accuracy rate and precision rate in Table A6. In Figure A19, I plot the one-time individual test results against the probabilities calculated based on Bayesian posterior distribution based on the formula above. The x-axis for the upper picture in Figure A19 is for the Probability of "Bad" Tier, and the x-axis for the lower picture in Figure A19 is for the Probability of "Good" Tier.

Table A6: Confusion Matrix - Added Water

Individual Test

Prediction	Good	Normal	Bad
Good	**436**	20	390
Normal	192	**35**	365
Bad	176	46	**1,898**

Notes: **Actual Numbers**

Individual Test

Prediction	Good	Normal	Bad
Good	**51.54%**		
Normal		**5.91%**	
Bad			**89.53%**

Notes: **Accuracy** (Row Percentage)

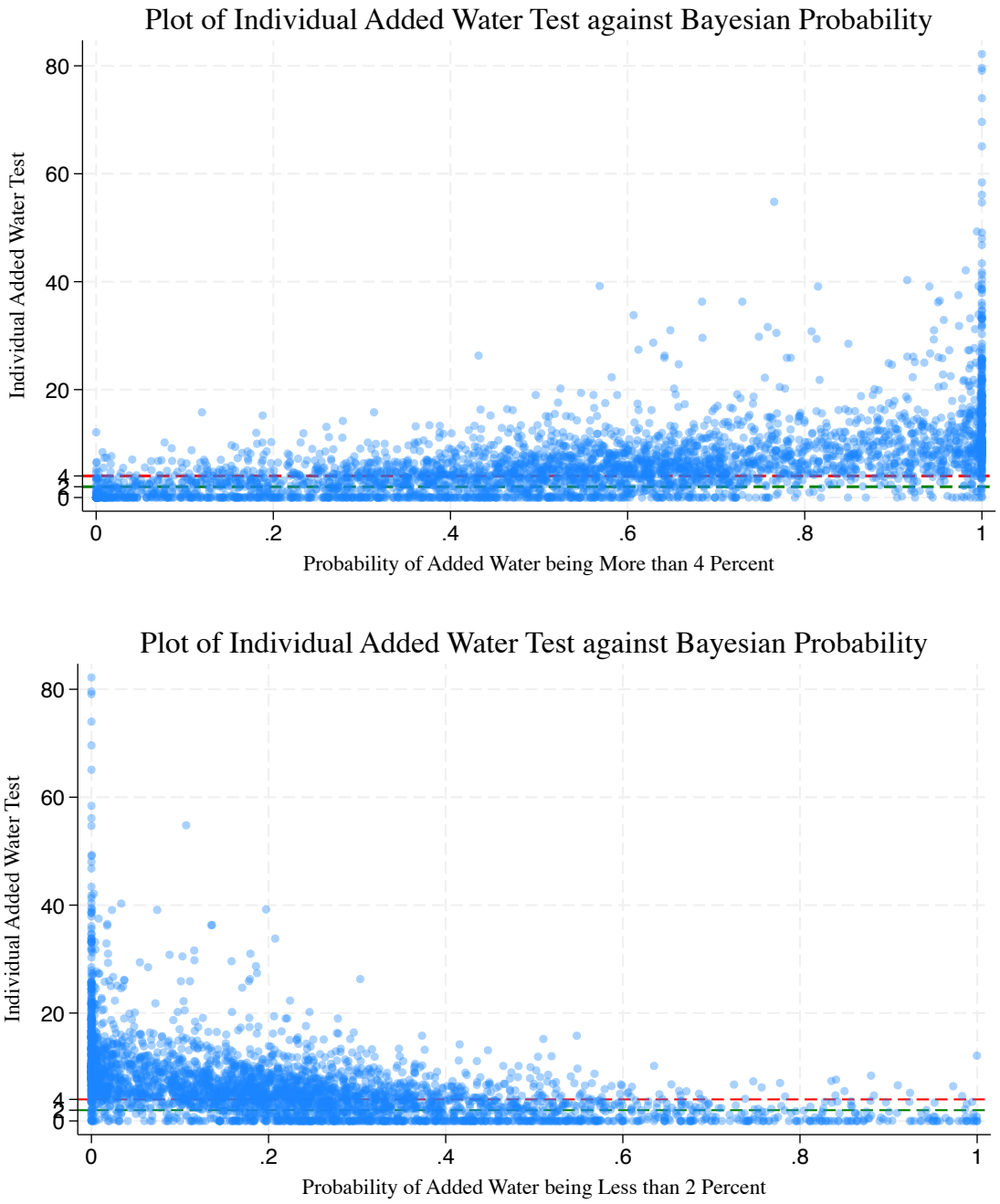
Individual Test

Prediction	Good	Normal	Bad
Good	**54.23%**		
Normal		**34.65%**	
Bad			**71.54%**

Notes: **Precision** (Column Percentage)

Notes: This table presents the confusion matrix for added water predictions, following the standard method for evaluating classification models in the field of machine learning. The upper panel displays the actual counts, the middle panel calculates accuracy using row percentages, and the bottom panel calculates precision using column percentages. It is important to keep in mind that farmers' milk quality is not set in stone, and one-time tests are not golden benchmarks to test the model's performance. As is discussed in section 2.2, one-time tests can only reflect the perfect information for the day when farmers' milk samples have been taken and tested, but not on other days. In contrast, Bayesian models on aggregated can-level do not give perfect information for any single day. They provide information about overall performance, but they could provide noisy signals.

Figure A19: Probability of High and Low Added Water



Notes: This figure illustrates the relationship between one-time individual test results and the probabilities derived from the Bayesian posterior distribution, calculated using the formula in Section C.2.2 for added water. Note that the upper and lower panels are not flips of each other, as there is also a probability for the "Normal" category, which represents added water levels between 2% and 4%.

For the interpretation of Figure A19 and Table A6, it is important to keep in mind that farmers' milk quality is not set in stone, and one-time tests are not golden benchmarks to test the model performance. As is discussed in section 2.2, one-time tests can only reflect the perfect information for the day when farmers' milk samples have been taken and tested, but not on other days. In contrast, Bayesian models on aggregated can-level do not give perfect information for any single day. They provide information about overall performance, but they could provide noisy signals.

In fact, another way to test model performance is to compare the expected number(from the model)[35] of farmers whose milk fails the quality standards if showing up on random days to test with the number of farmers who actually failed the quality standards based on that random sampling and individual tests. I did the exercise at baseline, and the model predicts 449 out of 940 farmers would fail the test and the individual tests have found 444 farmers who actually failed the quality standards, which is very close to the model prediction.

C.3 Model Setup for Butterfat

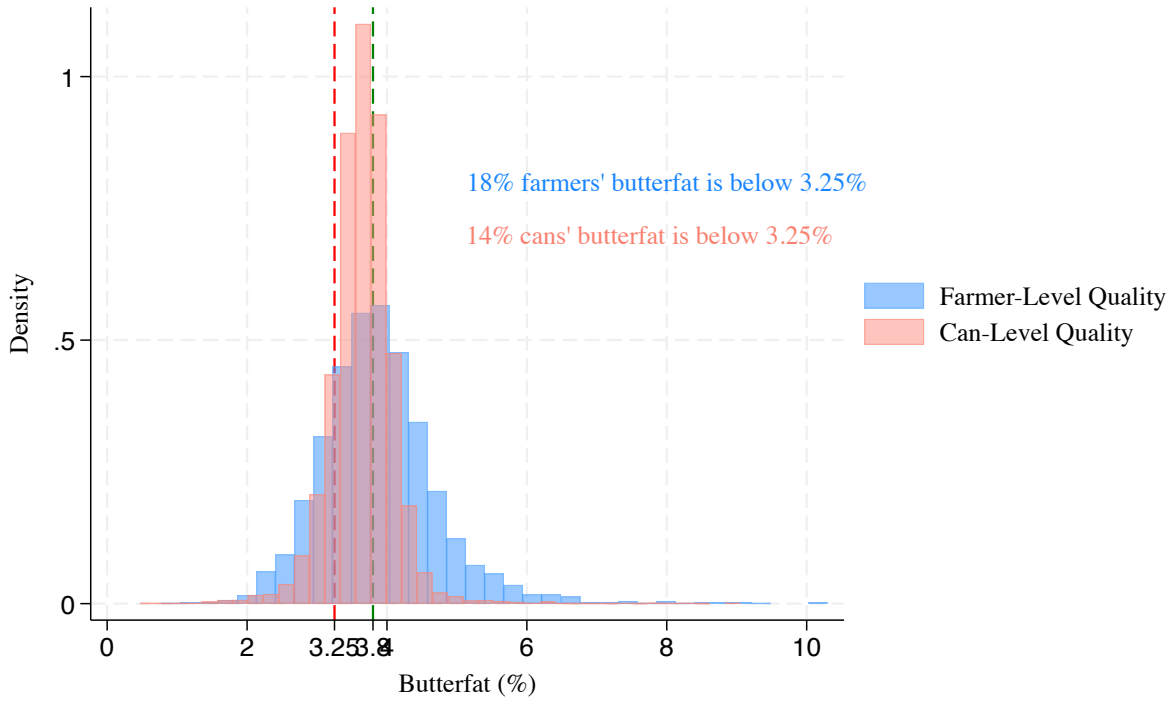
C.3.1 Priors for Butterfat

The model for butterfat is almost the same, except that the priors for butterfat and added water are different. Unlike added water, which is a non-negative value, butterfat is a strictly positive value unless the milk is 100% pure water, so I changed the priors for butterfat to capture this.

Let us keep the notation of the quality parameter the same, and the distribution of the butterfat for both individual levels and can levels are shown in Figure A20:

[35] Summing up the Probability of "Bad" Tier across all individual i.

Figure A20: Butterfat Distribution (Pooled All Rounds)



Notes: This figure shows the distributions of butterfat at both the farmer and can levels. The butterfat is strictly positive, and its distribution at the farmer level aligns well with a Gamma distribution, which I have chosen as the prior for this parameter. The national KeBS standards require that butterfat must exceed 3.25%.

The individual butterfat level, ρ_i^{indiv}, is strictly positive, and based on the histogram of the observed data in Figure A20, I model the potential distribution of butterfat using the gamma distribution:

$$\rho_i^{indiv} \sim \text{Gamma}\,(\alpha_{\text{prior}}, \beta_{\text{prior}})$$

The PDF is given by:

$$P(\rho_i^{indiv}) = \begin{cases} \frac{\beta^\alpha}{\Gamma(\alpha)} (\rho_i^{indiv})^{\alpha-1} \exp(-\beta \rho_i^{indiv}), & \rho_i^{indiv} > 0 \\ 0, & \text{otherwise,} \end{cases}$$

where α denotes the shape parameter, and β is the rate parameter; The gamma function is defined as:

$$\Gamma(\alpha) = \int_0^\infty y^{\alpha-1} e^{-y} dy$$

Since there are two parameters, I fix the α to be the same as the shape parameter fitted on

the gamma distribution using pilot data, and then the rate parameter β is chosen to be an inverse gamma distribution: $\beta \sim \text{InvGamma}(3, 1)$.

C.3.2 Likelihood for Butterfat

I modeled the likelihood of butterfat the same as I modeled for added water in section C.1.2:

$$L\left(\mathbf{Y}^{data} \mid \rho_1^{indiv}, \rho_2^{indiv}, \ldots, \rho_n^{indiv}\right) = \prod_{j=1}^{J} L\left(y_j^{data} \mid \rho_1^{indiv}, \rho_2^{indiv}, \ldots, \rho_n^{indiv}\right)$$

$$= \prod_{j=1}^{J} \frac{\frac{1}{\sigma^{can}\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y_j^{data} - \rho_j^{can}}{\sigma^{can}}\right)^2}}{1 - \Phi\left(\frac{-\rho_j^{can}}{\sigma^{can}}\right)} = \prod_{j=1}^{J} \frac{\frac{1}{\sigma^{can}\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y_j^{data} - \sum_{i=1}^{m_j} w_{ij}^{data} \rho_i^{indiv}}{\sigma^{can}}\right)^2}}{1 - \Phi\left(\frac{-\sum_{i=1}^{m_j} w_{ij}^{data} \rho_i^{indiv}}{\sigma^{can}}\right)},$$

σ^{can} is assigned an inverse gamma prior:

$$\sigma^{can} \sim \text{InvGamma}(3, 1)$$

C.3.3 Posteriors for Butterfat

Given the above setup, Bayes' theorem gives:

$$P\left(\rho_1^{indiv}, \rho_2^{indiv}, \ldots, \rho_n^{indiv} \mid \mathbf{Y}^{data}\right) \propto$$

$$L\left(\mathbf{Y}^{data} \mid \rho_1^{indiv}, \rho_2^{indiv}, \ldots, \rho_n^{indiv}\right) \times P\left(\rho_1^{indiv}\right) \times P\left(\rho_2^{indiv}\right) \times \ldots \times P\left(\rho_n^{indiv}\right)$$

$$= \prod_{j=1}^{J} L\left(y_j^{data} \mid \rho_1^{indiv}, \rho_2^{indiv}, \ldots, \rho_n^{indiv}\right) \times \prod_{i=1}^{n} P\left(\rho_i^{indiv}\right)$$

$$= \prod_{j=1}^{J} \frac{\frac{1}{\sigma^{can}\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y_j^{data} - \rho_j^{can}}{\sigma^{can}}\right)^2}}{1 - \Phi\left(\frac{-\rho_j^{can}}{\sigma^{can}}\right)} \times \prod_{i=1}^{n} \frac{\beta^{\alpha}}{\Gamma(\alpha)} (\rho_i^{indiv})^{\alpha-1} \exp(-\beta \rho_i^{indiv})$$

$$= \prod_{j=1}^{J} \frac{\frac{1}{\sigma^{can}\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y_j^{data} - \sum_{i=1}^{m_j} w_{ij}^{data} \rho_i^{indiv}}{\sigma^{can}}\right)^2}}{1 - \Phi\left(\frac{-\sum_{i=1}^{m_j} w_{ij}^{data} \rho_i^{indiv}}{\sigma^{can}}\right)} \times \prod_{i=1}^{n} \frac{\beta^{\alpha}}{\Gamma(\alpha)} (\rho_i^{indiv})^{\alpha-1} \exp(-\beta \rho_i^{indiv})$$

For the posterior full conditional distribution of each ρ_i^{indiv}, we have the following:

$$P\left(\rho_i^{indiv} \mid \rho_{-i}^{indiv}, \mathbf{Y}^{data}\right) \propto P\left(\rho_1^{indiv}, \rho_2^{indiv}, \ldots, \rho_n^{indiv} \mid \mathbf{Y}^{data}\right)$$

$$\propto \prod_{j=1}^{J} \frac{\frac{1}{\sigma^{can}\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y_j^{data} - \sum_{i=1}^{m_j} w_{ij}^{data} \rho_i^{indiv}}{\sigma^{can}}\right)^2}}{1 - \Phi\left(\frac{-\sum_{i=1}^{m_j} w_{ij}^{data} \rho_i^{indiv}}{\sigma^{can}}\right)} \times \prod_{i=1}^{n} \frac{\beta^{\alpha}}{\Gamma(\alpha)} (\rho_i^{indiv})^{\alpha-1} \exp(-\beta \rho_i^{indiv})$$

$$\propto \prod_{j=p}^{k} \frac{e^{-\frac{1}{2}\left(\frac{y_j^{data} - \sum_{i=1}^{m_j} w_{ij}^{data} \rho_i^{indiv}}{\sigma^{can}}\right)^2}}{1 - \Phi\left(\frac{-\sum_{i=1}^{m_j} w_{ij}^{data} \rho_i^{indiv}}{\sigma^{can}}\right)} \times (\rho_i^{indiv})^{\alpha-1} \exp(-\beta \rho_i^{indiv})$$

$$\propto \frac{\exp\left\{-\frac{1}{2}\left[\sum_{j=p}^{k}\left(y_j^{data} - \sum_{i=1}^{m_j} w_{ij}^{data} \rho_i^{indiv}\right)^2 + \left(2\beta \rho_i^{indiv}\right)\right]\right\} \times (\rho_i^{indiv})^{\alpha-1}}{\prod_{j=p}^{k}\left[1 - \Phi\left(\frac{-\sum_{i=1}^{m_j} w_{ij}^{data} \rho_i^{indiv}}{\sigma^{can}}\right)\right]},$$

where p and k mean that individual farmer i show up in can p, can $p+1$, ..., can k. Note that any individual i showed up at least one time, and individual $i's$ milk was divided into $(k+1)$ different aggregated containers on that single day. We also have $\Phi\left(\frac{-\sum_{i=1}^{m_j} w_{ij}^{data} \rho_i^{indiv}}{\sigma^{can}}\right) = \int_{-\infty}^{\left(\frac{-\sum_{i=1}^{m_j} w_{ij}^{data} \rho_i^{indiv}}{\sigma^{can}}\right)} \phi(u)du$, $\phi(u)$ is the pdf of standard normal distribution: $\phi(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2}$.

C.4 Estimation and Model Performance for Butterfat

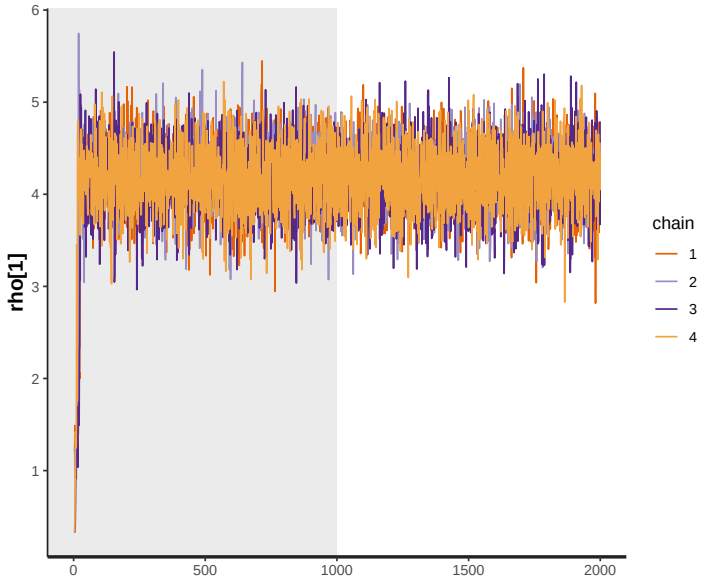
C.4.1 Estimation - Hamiltonian Monte Carlo (HMC) for Butterfat

I obtained point estimates and credible intervals for each ρ_i^{indiv}, β, and σ^{can}. I attach the model estimation based on the modified Hamiltonian Monte Carlo (HMC) simulations[36] in Figure A21 and compare the Bayesian point estimates(posterior mean) to the Individual test results among 940 dairy farmers from cooperatives in two different counties in Figure A22.

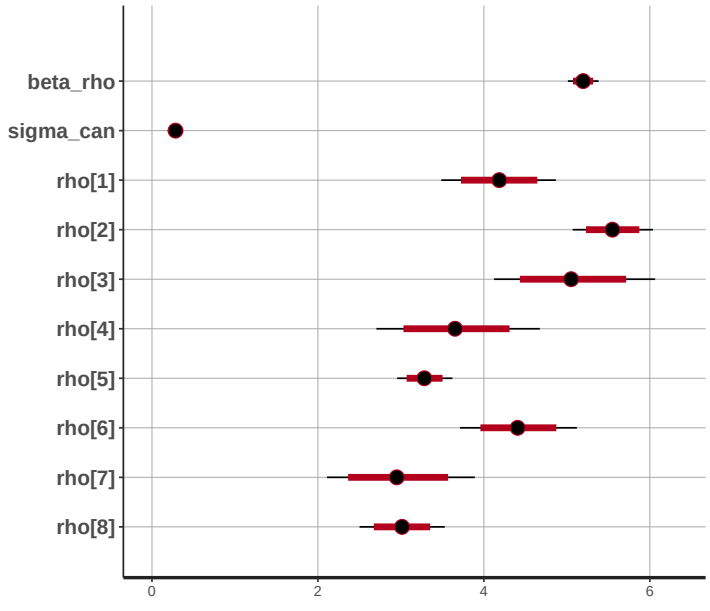
[36] Specifically, No-U-Turn sampler.

Figure A21: Hamiltonian Monte Carlo (HMC) Simulations for Butterfat Estimation

(a) Panel A: Trace Plot

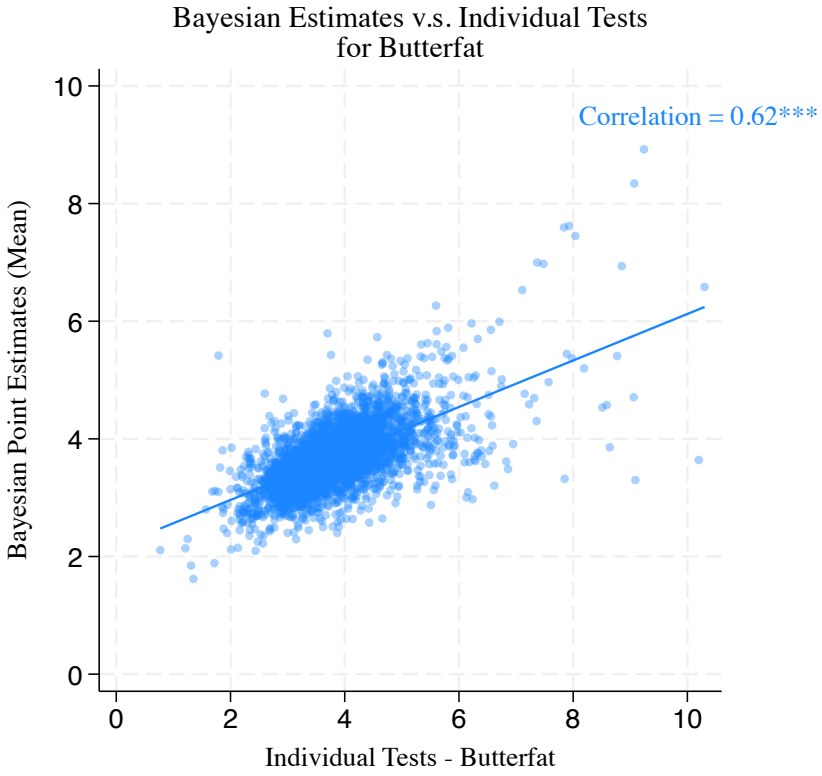


(b) Panel B: Caterpillar Plot



Notes: This figure presents the model estimation results based on the modified Hamiltonian Monte Carlo (HMC) simulations. Panel A displays the trace plots of the Markov chains, illustrating the time series of posterior draws. The first 1,000 draws, shaded in gray, represent the warmup phase, while the subsequent 1,000 draws from the post-warmup phase are used to approximate the posterior distribution of the parameter. Four independent chains were used, and their mixing was verified to ensure proper convergence during the post-warmup phase. Panel B presents the point estimates (black dots) alongside the posterior credible intervals, with the inner interval representing 80% credibility and the outer interval representing 95%, to illustrate the uncertainty in the estimates.

Figure A22: Comparison between Model Prediction and Individual Tests on Butterfat



Notes: This figure compares Bayesian point estimates of average butterfat in the milk over the monitoring period with random individual test results. Both the X-axis and Y-axis represent absolute quality levels measured as percentages. The Pearson correlation coefficients between the Bayesian point estimates and the individual test results are calculated, pooling data across all four rounds of quality monitoring. ***, **, and * represent significance at 1%, 5%, and 10% respectively.

C.4.2 Posterior Probability Classification for Butterfat

Likewise, I use the entire posterior distribution to classify farmers into three categories based on the minimally required standards posted by the government (KEBS) and the processors' quality rewarding standards for butterfat. The probabilities for each tier[37] are calculated as follows: - Probability of "Bad" Tier: $P_{\text{Bad}}(i) = \int_0^{\theta_{\text{low}}} f\left(\rho_i^{indiv}\right) d\rho_i^{indiv}$

- Probability of "Normal" Tier: $P_{\text{Normal}}(i) = \int_{\theta_{\text{low}}}^{\theta_{\text{high}}} f\left(\rho_i^{indiv}\right) d\rho_i^{indiv}$

- Probability of "Good" Tier: $P_{\text{Good}}(i) = \int_{\theta_{\text{high}}}^{\infty} f\left(\rho_i^{indiv}\right) d\rho_i^{indiv}$,

[37]Note that for butterfat, the higher, the better.

where $\theta_{low} = 3.25\%$, which is the minimal required standards by the KeBS; $\theta_{high} = 3.8\%$, which is the processor's standard for quality rewarding.

I compare the predicted categories and categories based on a one-time test in the confusion matrix below and calculate the accuracy rate and precision rate in Table A7. Again, it is important to keep in mind that farmers' milk quality is not set in stone, and one-time tests are not golden benchmarks to test the model performance. In Figure A23, I plot the one-time individual test results against the probabilities calculated based on Bayesian posterior distribution based on the formula above. The x-axis for the upper picture in Figure A23 is for the Probability of "Bad" Tier, and the x-axis for the lower picture in Figure A23 is for the Probability of "Good" Tier.

Table A7: Confusion Matrix - Butterfat

Individual Test

Prediction	Good	Normal	Bad
Good	**1,302**	327	104
Normal	358	**447**	179
Bad	181	227	**433**

Notes: **Actual Numbers**

Individual Test

Prediction	Good	Normal	Bad
Good	**75.13%**		
Normal		**45.43%**	
Bad			**51.49%**

Notes: **Accuracy** (Row Percentage)

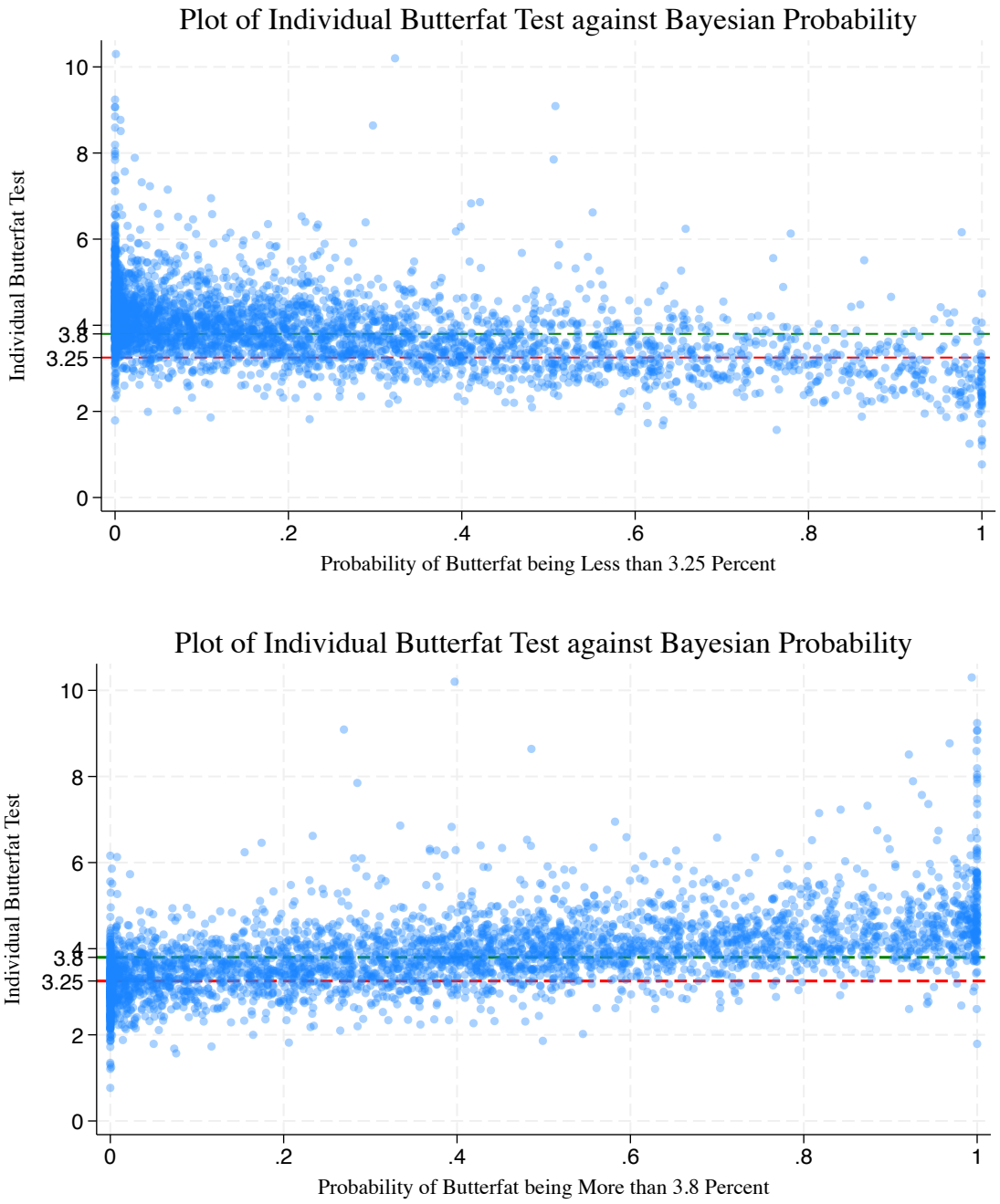
Individual Test

Prediction	Good	Normal	Bad
Good	**70.72%**		
Normal		**44.66%**	
Bad			**60.47%**

Notes: **Precision** (Column Percentage)

Notes: This table presents the confusion matrix for butterfat predictions, following the standard method for evaluating classification models in the field of machine learning. The upper panel displays the actual counts, the middle panel calculates accuracy using row percentages, and the bottom panel calculates precision using column percentages. It is important to keep in mind that farmers' milk quality is not set in stone, and one-time tests are not golden benchmarks to test the model's performance. As is discussed in section 2.2, one-time tests can only reflect the perfect information for the day when farmers' milk samples have been taken and tested, but not on other days. In contrast, Bayesian models on aggregated can-level do not give perfect information for any single day. They provide information about overall performance, but they could provide noisy signals.

Figure A23: Probability of High and Low Butterfat



Notes: This figure illustrates the relationship between one-time individual test results and the probabilities derived from the Bayesian posterior distribution, calculated using the formula in Section C.4.2 for butterfat. Note that the upper and lower panels are not flips of each other, as there is also a probability for the "Normal" category, which represents butterfat levels between 3.25% and 3.8%.