Reviewer #1

1. The idea to use machine learning to create a vFI is novel and interesting. While the authors have done a thorough job describing their approach, It is not entirely clear how this instrument will be operationalized. This is important because the authors argue that the vFI will enhance the utility of the FI for mouse evaluations. If it is tricky to operationalize this could adversely affect its uptake by the research community. Can the authors make this vFI easy to implement?

2. The comments made about the inter-rater reliability of the manual mouse FI are somewhat more pessimistic than the literature would suggest. Papers that have addressed this issue have used different raters who rate the same mice. This appears to contrast with what the authors have done to assess reliability. These previous studies have also compared results with interrater reliability statistics (e.g. intra-class correlation coefficients and Cohen’s kappa statistic). Together (e.g. PMID: 25205762; PMID: 25711530; PMID: 28463656; PMID: 25838548), they show very good agreement between raters, although one study did note that agreement between scientist raters and animal care technician raters was not high (PMID: 25838548). Of particular relevance to the submitted paper, an earlier study found excellent agreement between experienced and inexperienced scientist raters (PMID: 28463656). The authors need to revise  
their discussion of previous work on reliability to reflect the published literature more clearly. They also need to describe how they assessed reliability (unclear from Figure 1C). Did they, in fact, have different raters assess the same mice? This is not clear from the methods section or the text at present.

We agree with the reviewer that our framing of the problem of inter-rater reliability was more negative than the literature suggests and so we have revised our text to reflect that. However, the above papers (PMID: 25205762: PMID: 25711530; PMID: 28463656; PMID: 25838548) talk about the importance of discussion and refinement to achieve very high/excellent inter-rater reliability. Feridooni et al., (2015) (PMID: 25205762) found initial disagreement between scorers greater than 25% on 11 deficits. They were able to greatly resolve this disagreement over the next 2 rounds by comparing rating procedures and expanding descriptions of the evaluation. Kane et al., (2015) found that many of the same deficits as Feridooni et al. showed the most disagreement between scorers and noted that these deficits tend to require some degree of subjective judgement in their scoring, which may be resolved with discussion. Kane et al., (2015, 2017) (PMID: 28463656; PMID: 25838548) find that without discussion and refinement, inter-rater reliability does not improve solely by practice and experience. All together this suggests that the subjectivity of scoring certain deficits may lead to disagreement on those deficits between labs which may not communicate. Inter-lab differences in frailty scoring may also have factors not found in intra-lab studies.

We did not have the same mouse tested by different scorers in this version of the paper; four different scorers were used to test different batches of mice. We visually saw a pattern in the data by scorer so we (Gautum explain this part)

3. Likewise, if the sentence “They were tested in 2 rounds approximately 5 months apart” (Methods section 5.1) helps to strengthen generalizability, replication in a different sample would be beneficial, especially if the system readily allows high throughput. Does this improve the information of the 20/80 split design here? It is also unclear from the report whether the authors have repeated observations in the same animals 5 months apart. Similarly, it is not clear from the present data whether the vFI replaces or complements the traditional FI in explaining outcomes. This question needs to be tested to support the claim that a vFI is all that is needed for high throughput Characterization of frailty in ageing animals.

(I’m not sure I understand the first part of this comment. )

117 of the mice from the first round of testing were repeated in the second round of testing 5 months later. To aid in clarity, we have added a table of all the mice and when they were tested.

As for the next point, this is an important consideration. Currently, we imagine that for labs, especially smaller labs, who have not used the FI in the past and do not have a dedicated frailty expert, the vFI could be a sufficient tool to do frailty research. For labs who already competently use FIs, the vFI could be used to complement the manual FI by adding rich information on changes in behavior and physiology for analysis. We believe that our error is small enough that it justifies replacing the manual FI in certain contexts where it is needed: where throughput is of high concern or where there is lack of expertise in FI.

4. The utility/value of the vFI would be more convincing if the authors could demonstrate that it was responsive to an intervention. Can the authors demonstrate that the vFI responds to interventions that make frailty better or worse?

An interventional experiment would be a great extension of this work and we plan to do this in the future; we currently are running an experiment on known dietary interventions on frailty. Unfortunately, this is unfeasible to include in this paper due to time and scope.

5. The authors report some sex differences in FI scores (Figure 3A), with female mice having lower FI scores than males at many ages. They also indicate that very little is known about sex differences in frailty. In fact, quite a few studies have addressed this question. Most report that female C57Bl/6 mice have higher FI scores than males (a few examples are PMID: 28325885; PMID: 29788087; PMID: 24051346) although some report no sex difference (e.g. PMID: 29924918). A study of sex differences in the frailty phenotype also shows higher frailty in females (PMID: 31355774). Especially given the comments by the authors that most studies of frailty in mice have used males, the paper needs to be revised to reflect the current literature. Literature on sex differences in frailty should be cited and the authors should discuss why their results differ from several previous reports. 

We agree that in light of recent research it may be inaccurate to say that sex differences in frailty are not well studied, so we have amended that statement. However, it may still be accurate to say sex differences in frailty *indexing* (specifically cFI) are still less studied, as much of the research has been on sex differences has been on other frailty measures like frailty phenotyping or, like in PMID: 28325885, in other forms of frailty quantification. Given that it has been found that in mice, like humans, frailty phenotyping and cFIs do not always identify the same individuals as frail (PMID: 28549083 finds only 50% agreement in the mice classified as frail, comparable to the agreement in human studies), it is unclear whether other forms of frailty quantification are directly comparable to the manual FI. Nevertheless, we have modified our text to reflect the broader research and removed the statement about the underrepresentation of female mice in FI research.

We have also amended our statement about the evidence on the sex-frailty paradox in mice to reflect the sources cited by the reviewer.

6.The work is generally well written in these sections. However, the vFI does not assess a wide range of body systems (largely gait, mobility, body composition and grooming). Many other body systems are included in the manual FI. This limitation of the vFI should be addressed in the discussion.

It is definitely true that the vFI does not specifically asses the range of body systems that is scored in the manual FI. Indeed we would very much like to expand the types of features we include in our vFI. However, we believe that the vast amount of sensitive measures we take (ex: 24 different measures of gait alone) contains implicit information about many body systems. It is easy to imagine how manual FI items like vision loss, vestibular disturbances, tremor, tail stiffening, and others would influence gait and behavior. Complex behaviors in the open field require the combined input of a variety of body systems and cognition. Again though, in the future, we would like to expand our features to make the inclusion of other body systems more explicit. For example, coat changes that can be seen in the video could be used as a feature.

Reviewer #2

1. The authors provide an impressive dataset that is of high value for the field. This dataset contains OFT recordings of almost 500 animals with corresponding manual FI values (an impressive feat). However, in the current manuscript the data (video recordings, feature data for each individual animal, manual FI score for individual animals, predicted age/FI scores for individual animals) and code is not made accessible to reviewers. Only a statement that it will be made public can be found. Therefore, only superficial feedback about the machine learning validity can be given at this point. A revised manuscript should include data that can be independently validated.

2. Based on what is presented we are skeptical that the predicted vFI score is not just predicting age rather than the FI value. Age and FI are correlated to each other (Figure 1B) and to the same features (Figure 4B and 4D are remarkably similar). Figure 4 J shows that in cases where the manual FI score (blue dots) deviate far from the actual age (red dot in Figure 4I at the same index) the prediction (gray dot in Figure 4J) lies much closer to FI values of animals of similar age, rather than to animals of similar FI scores. This indicates that the model maybe looks for age specific cues to decide what FI values to predict, rather than finding FI specific cues. A clear improvement of the random forest regression model using video feature data over a FI prediction based on age alone (linear or polynomial regression model) needs to be rigorously demonstrated. If age alone is a similarly good predictor using a vFI is not very useful, since age of experimental animals is a readily  
available information in any animal facility.  
  
3. If point 2) can be addressed, the authors should further investigate the difference between the age predictor and the vFI predictor. The feature importance analyses already used in the manuscript are an excellent tool for this; however, they need to be repeated for the age predictor and presented and compared side-to-side to the vFI predictor. This should ideally illustrate features or feature pairs which are highly informative for FI prediction, but not for age prediction and vice versa.  
  
4. The authors state that 95% of the actual values are within the prediction intervals (Figure 4 I and J). However, this is less impressive given that the prediction intervals span most of the data range. Unless prediction intervals can be improved, we would recommend removing them completely.

5. In the discussion the authors claim that they provide a tool that can be used for high throughput studies to assess vFI automatically. However, there are some important missing discussion points.  
  
a. The already trained models maybe only work accurately in a very limited context where a similar strain of mice in a similar set up are used.  
b. If an external researcher tries to adapt the system for their set up it might require the generation of new trainings data that could be very time intensive  
c. These issues could be addressed experimentally by the authors by:   
i. using annotated test data from an external collaboration (different mice / setup / recording) to benchmark their model.  
ii. Analyzing how scaling the test-train split influences the model’s performance. Maybe a smaller training set is already sufficient to reach a similar performance. In this case generating a new training sets would be less of a hurdle for a new setup

Reviewer #3

1. First, I am concerned with the age distribution and statistics underlying their performance. There is a strongly bimodal distribution of ages, shown in 4a, i, which predisposes the analyses to Simpson’s paradox. The model predictions appear to be even more strongly bimodal, and perform much more poorly over the intermittent ages between 50 and 100. In the attached figure I’ve color coded this so as to be abundantly clear. I’d be happy to clarify Simpson’s paradox further, but the correlation values may be suspect when most of your data come from one of two subgroups.  
Said a different way, it is very easy to determine a mouse to be young/old. I’m guessing with a single measure, I could simply threshold and determine 50 or 100 weeks. In the data shown, the difficulties of their model may be both due to the lack of training data on intermediate ages and because these intermediate ages are more difficult. One would assume then, that the distribution of samples should be unimodal around ~75 weeks, not bimodally young and old.  
  
On the other hand, the range is bounded oddly for an aging study. As a comparison, the authors cite Schultz et al., a majority of the data come from animals over 120 weeks (28 months x 4.3 weeks/month). Rockwood has similar distributions. Here, only 1? Maybe 7? of the nearly 150 datapoints are from mice in this range, and the model itself appears incapable of scoring above this near-100 week level, nor of an FI score above 6-ish. This is a problem. I would strongly recommend expansion of the training data/improvement of the model, as well as applicability (claiming 148 weeks with an n of 1 for anything near that value is sub-ideal). The authors' 95% percenticle upper age is approx 23 months, or the upper bound of Rockwell’s "Middle age” classification for mice (not “Older”)  
Moreover, a comparison with a similar measure for age prediction, e.g. FRIGHT would be very, very useful for quantifying the power of the method.   
  
Somewhat related to the above: I apologize, I do not understand how this works. All of the FI’s in 4j are under 11, and 95% of them are between 1 and 9. The histogram in 4c seems to mirror this, although there may be one? more animal with an FI of 13? Regardless, the FI range reported in the figures appears to be less than half of that mentioned in the text.   
"We can also predict the FI score to be within 1.03 ± 0.08 (3.9% ± 0.3%) of the actual frailty index.” also "The distribution of FI scores (range 0-27)”   
This has a demonstrable effect on how to interpret the ability to interptet the quality of the model.  
  
2. While the computational approaches appear rigorous, and I applaud the Kumar group for their series of works within a common theme, the work faces a challenge with impact: it is unclear to me how another group - particularly one with less computational skill - would be able to apply these methods to their own research. After all, a tool is only as valuable as its ability to be wielded. The utility of this paper to the field would be greatly increased if this RF was made both accessible and feasibly generalized. I could see this approach becoming the go-to for the field, and an open-souce approach is highlighted well in an open-source journal. This may be the intention of the authors, and if so, I would advise them to focus much more of the text to this. Considerable documentation is provided for the acquisition of the information. In my opinion, about half of this could be shifted to the supplement, especially if it made room for a figure highlighting the integrative use of  
the tool. A huge, well-trained dataset like this is unlikely to be accessible to many other groups, and thus my opinion for this work fitting Nature Aging - rather than a methods journal - is heavily swayed by this. Again, one of the fantastic parts of this study is the potential to remove inter-rater, and indeed inter-lab variability, but only if it can be applied.