Cell Genomics Transparent Peer Review Record

Machine learning enables new insights into genetic contributions to liver fat accumulation

Mary E. Haas,[1,2]* James P. Pirruccello,[1,4,5,6]* Samuel N. Friedman,[6]* Minxian Wang,[1] Connor A. Emdin,[1,4] Veeral H. Ajmera,[7] Tracey G. Simon,[4,8] Julian R. Homburger,[9] Xiuqing Guo,[11] Matthew Budoff,[11] Kathleen E. Corey,[4,8] Alicia Y. Zhou,[9] Anthony Philippakis,[6,10] Patrick T. Ellinor,[1,4,5,6] Rohit Loomba,[7] Puneet Batra,[6] Amit V. Khera[1,3,4,5,6]

Summary

Scientific Editor:	Orli Bahcall
Initial submission:	1/12/2021
Revision received:	7/13/2021
Accepted:	9/09/2021
Rounds of review:	2
Number of reviewers:	3

Referee reports, first round of review

Reviewer #1: This study is a well performed and well described assessment of liver fat, based on the recently available liver imaging data in the UK Biobank. The authors have addressed all the comments from the previous journal very thoroughly. I have only minor comments:

1. To address the issue of representativeness , the authors could include the baseline summary details of the UK Biobank participants who did not participate in the imaging study (although this may be complicated by those who have not yet been invited, or invited but not yet imaged. From the BMIs, the individuals look a little healthier than overall UK Biobank.
2. Lines 170 to 183. Comparing imaging to standard clinical biomarkers . BMI alone has a correlation of 0.48 with imaged liver fat, but a model of 25 variables, including circulating markers of liver fat , has only a correlation of 0.46. This seems very odd. How about using the markers of liver fat alone, especially those that are more specific to NAFLD rather than alcohol ?

Reviewer #2: Our comments:
It was a pleasure to read this paper which is a major advance. The authors should be congratulated. We have very little to add as the first round of extensive reviews identified the really important information that was missing.

Minor comments:
The cohort is 97% white/European so the opening sentence of the Discussion really needs to emphasize that this is really a white, middle aged cohort. This is left to a throw away sentence in the final paragraph.

"Rare inactivating variants" should be defined first time this nomenclature is used (Who was considered to be a rare inactivating variant carrier was not defined before p.18, line 954 as part of the Methods).

Several subsets of UK biobank participants were studied. It was not clear if principal components (PCs) were calculated based on all genotyped UK biobank participants, or only on individuals with european ancestry, or in the subsets investigated? Furthermore, we are wondering if there was a rationale why only the first 5 PCs were included in the polygenic score analysis of UK Biobank participants, as compared to the CVAS and association analysis of top CVAS variants with blood biomarkers/physician diagnosis performed in UK Biobank. In the latter analyses, the first 10 PCs were included in the statistical models. In the polygenic score analysis you show that the polygenic score improves discrimination from 0.55 to 0.59, but this is the analysis where you only include the first 5 PCs. Why did you not include the 10 first as in the CWAS?

The authors used the definition of excessive alcohol intake based on UK and US guidelines to calculate the prevalence of excessive alcohol intake within the UK Biobank. Inconsistently to this, the two sensitivity analyses (described in line 293 onward) were only performed using US cut-offs, and not using the more stringent UK cut-offs. It would be interesting to run a GWAS excluding individuals based on the UK cut-off for excessive alcohol consumption to check whether the results for the 8 variants change.

Did you exclude individuals based on other diseases such as hepatitis b and hepatitis c infections?

On page 12, line 317 the authors analyze the variants identified by the GWAS using liver biomarkers and clinical diagnoses.The biomarkers were generally obtained ~10 years prior to the imaging. Are these liver biomarkers in general stable or do they fluctuate? It could be mentioned that these biomarkers are obtained 10 years earlier than the images so the reader is aware when interpreting the results.

In the "teacher model" (line 702), the individuals had a full set of 10 standard images. How many images are available for the remaining participants?

The readers may be curious as to how liver fat is normally calculated in the clinic when radiologists read out liver fat. And how different is that that different from the machine learning algorithm that you apply? In many ways an automated machine-learning algorithm has obvious advantages.

Reviewer #3: This is a very interesting paper that integrates use of several methodological tools to conduct an end-to-end analysis of clinical and genetic contributions to liver fat accumulation.

As requested by the editor, I focused my review on the machine learning aspects of the manuscript. Especially with the additions that were made per Reviewer 1's request, I feel that the machine learning methods are sound and are described to a sufficient level of detail. The training and evaluation framework is a bit complicated due to the need to use a teacher/learner approach but it is appropriate for the data that was available.

Major comments:

- The first heading of results is, "Machine learning model enables near-perfect quantification of hepatic fat based on raw abdominal MRI imaging". However, the evaluation measure reported is Pearson correlation, which does not exactly capture quantification in the sense of measuring how close the predicted liver fat is to the true liver fat, since Pearson correlation is invariant to scaling/shifting, so it only captures that there is a linear relationship between the true and predicted liver fat values but not that the values are actually close to each other.

A better measure would be something like mean absolute error, which does measure how close the true and predicted values are to each other. Based on Figure S2, it appears that the machine learning model will still have strong performance relative to a true quantification measure like mean absolute error, but this should be reported.

Note, this difference is of practical and not just terminological importance, since the predicted liver fat values themselves are used in downstream analysis, so the actual number being correct is quite important, rather than just the correlation.

Minor comments:

- In Table 1, very small P values (like 1x10-92 and 2x10-234) are reported. This level of precision in differences of clinical characteristics is not meaningful even with 30k patients. I would recommend truncating at a threshold like P<0.001 as is conventional when reporting differences of clinical characteristics.

Reviewer #4: Comments enter in this field will be shared with the author; your identity will remain anonymous.In the paper "Machine learning enables new insights into clinical significance of and genetic contributions to liver fat accumulation", the authors used a 2D CNNs to estimate liver fat percentage from abdominal MRI. The model was trained on a small set of labeled data and can estimate liver fat percentage for a large set of unlabeled data and therefore can increase the sample size for association studies and identify additional genetic variants. The follow-up investigation is quite detailed and they have responded other reviewers' comments. Overall, this study is well designed, and the manuscript is well written. I only have two concerns.

1. Regarding the evaluation of 2D CNN model. The authors claimed that the CNN model produced correlation coefficients of 0.97 and 0.99 on hold-out testing sets, and therefore the estimated liver fat percentage for the unlabeled sampled is accurate for downstream association analysis. Usually for evaluating a supervised ML model, we divide the samples into a number of folds and use cross-validation to evaluate. For each evaluation, a train set is used to train the model, and a tune set is used to choose the parameter and a test set is used to evaluate the performance. If 10-fold cross valuation is used, the performance on the test set from all the 10 times are reported. I wonder whether the authors can use a careful 10-fold cross-validation to evaluate the correlation coefficients? From what the authors described, they only used a train/test split to evaluate, and the performance can depend on the parameter choices.

2. Even that the CNN model is near-perfect accurate, it is not 100% accurate, and I wonder whether it will affect the downstream association results. I wonder whether the CNN model produces any confidence output for each sample. If yes, is it possible we eliminate a small portion of samples whose liver fat percentage estimation from the model is not confident, and then check whether the association results remain the same? Will the 8 variants remain the top 8 on the list? If this is possible, I would suggest you try a number of portions (e.g. 1%, 5% and 10%).

Author response to the first round of review

Reviewer #1 Summary:

Author Reply: We appreciate and agree with the summary above.

Reviewer #1 Minor Comment 1:

Author Reply: We agree that these additional details are of interest.

Manuscript Change(s): In response, we have included the baseline characteristics of UK Biobank participants stratified by inclusion in the imaging substudy in Table S1:

Table S1. Baseline characteristics of participants in UK Biobank stratified by inclusion in the imaging substudy, Related to Table 1.

	Overall	Imaged	Not Imaged	P-value
	(N=502521)	(N=36703)	(N=465818)	
Female	273394 (54.4%)	19049 (51.9%)	254345 (54.6%)	<0.001

Age at enrollment, years	56.5 (8.10)	54.9 (7.47)	56.7 (8.13)	<0.001
Age at imaging, years	NA	64.2 (7.56)	NA	NA
Self-reported ethnicity				
White	472711 (94.1%)	35572 (96.9%)	437139 (93.8%)	<0.001
Black	8034 (1.6%)	214 (0.6%)	7820 (1.7%)	<0.001
Other Asian	3389 (0.7%)	165 (0.4%)	3224 (0.7%)	<0.001
South Asian	8024 (1.6%)	313 (0.9%)	7711 (1.7%)	<0.001
Multiple, other or not provided	10363 (2.1%)	439 (1.2%)	9924 (2.1%)	<0.001
Coronary artery disease	17404 (3.5%)	1076 (2.9%)	16328 (3.5%)	<0.001
Diabetes	27848 (5.5%)	1808 (4.9%)	26040 (5.6%)	<0.001
Obese	122252 (24.3%)	6495 (17.7%)	115757 (24.9%)	<0.001
Hypertension	147343 (29.3%)	10289 (28.0%)	137054 (29.4%)	<0.001
Medications				
Anti-hypertensive therapy	104005 (20.7%)	4940 (13.5%)	99065 (21.3%)	<0.001
Lipid-lowering therapy	98894 (19.7%)	5552 (15.1%)	93342 (20.0%)	<0.001
Anthropometric data				
Weight, kg	78.1 (15.9)	76.8 (14.8)	78.2 (16.0)	<0.001
Waist-to-hip ratio	0.87 (0.09)	0.86 (0.09)	0.87 (0.09)	<0.001
Body-mass index, kg/m^2	27.4 (4.80)	26.6 (4.19)	27.5 (4.84)	<0.001
Body fat, %	31.5 (8.55)	30.0 (8.17)	31.6 (8.57)	<0.001
Estimated untreated systolic blood pressure, mmHg	141 (20.7)	137 (19.3)	141 (20.8)	<0.001
Alcohol consumption				
Weekly drinks, U.S. standard	4.84 (6.74)	5.48 (6.37)	4.79 (6.76)	<0.001
Weekly drinks, U.K. standard	8.47 (11.8)	9.58 (11.1)	8.38 (11.8)	<0.001
Excessive alcohol intake, U.S.	26408 (5.3%)	2015 (5.5%)	24393 (5.2%)	0.036
Excessive alcohol intake, U.K.	105842 (21.1%)	9066 (24.7%)	96776 (20.8%)	<0.001
Liver-associated biomarker concentrations				
Alanine aminotransferase, IU/L	23.5 (14.2)	23.0 (13.9)	23.6 (14.2)	<0.001
Aspartate aminotransferase, IU/L	26.2 (10.7)	25.8 (10.5)	26.3 (10.7)	<0.001
Gamma glutamyltransferase, IU/L	37.4 (42.1)	33.7 (33.9)	37.7 (42.7)	<0.001

Estimated untreated lipid concentrations				
Total cholesterol, mg/dL	228 (42.4)	227 (40.7)	228 (42.5)	<0.001
LDL cholesterol, mg/dL	146 (33.3)	144 (32.0)	146 (33.4)	<0.001
HDL cholesterol, mg/dL	56.0 (14.8)	57.0 (14.5)	55.9 (14.8)	<0.001
Triglycerides, mg/dL	135 [94-197]	126 [89-184]	136 [95-199]	<0.001
Glycemic biomarker concentrations				
Glycated hemoglobin, %	5.46 (0.620)	5.36 (0.475)	5.47 (0.629)	<0.001
Random glucose, mg/dL	92.3 (22.4)	89.9 (17.5)	92.5 (22.7)	<0.001

Values correspond to number (%), mean (standard deviation), or median [interquartile range]. P-values correspond to chi-squared test or Wilcoxon rank sum for categorical and continuous variables, respectively, for imaged compared to not imaged. Obesity was defined as body-mass index ≥ 30 kg/m^2 (NHLBI Expert Panel, 1998); excessive alcohol intake, U.S. was defined as alcohol intake exceeding American Association for the Study of Liver Disease guidelines for NAFLD definition (Chalasani et al., 2018); excessive alcohol intake, U.K. was defined as alcohol intake exceeding the UK Chief Medical Officers recommendations (Department of Health, 2016). Diseases were defined as prevalent at time of initial assessment. Estimated untreated lipid measures and blood pressure were according to previously described adjustments (Ehret et al., 2016; Patel et al., 2020). NA, Not applicable.

We also added information about the overall cohort in the revised STAR Methods section:

> *"The UK Biobank is a prospective cohort study that enrolled 502,617 individuals aged 40-69 years of age from across the United Kingdom between 2006 and 2010 (Sudlow et al., 2015). As part of the study protocol, a subset of individuals underwent detailed imaging including abdominal MRI (Littlejohns et al., 2020) between 2014 and 2019, an average of 9.3 years after enrollment visit.* ***Participants who underwent imaging tended to be healthier than those who did not, as reflected by lower rates of obesity, coronary artery disease, and diabetes (Table S1).****"*

Reviewer #1 Minor Comment 2:

Author Reply: We agree that additional clarification is needed. In the previous submission, the correlation of BMI and liver fat was 0.42 within the holdout testing dataset of 1,214 used in machine learning model validation but 0.48 in the full dataset of 36,703 individuals.

To enable a more direct comparison, we have now updated the analysis in two ways: (i) restrict all analysis to the testing dataset of 1,214 individuals; (ii) update anthropometric data included in the model to parameters measured on the day of imaging as opposed to study enrollment.

For the model including clinical factors in addition to BMI, we included a range of variables reflecting potential markers of liver fat and NAFLD:

BMI, waist circumference, hip circumference, total body fat mass, total body fat percent, age at baseline, sex, height, weight, trunk fat mass, trunk fat percent, WHR, LDL cholesterol, total cholesterol, HDL cholesterol, triglycerides, systolic blood pressure, alkaline phosphatase, alanine aminotransferase (ALT), aspartate aminotransferase (AST), ALT/AST, gamma glutamyltransferase, hemoglobin A1c, random glucose, and C-reactive protein

Using this expanded set of clinical factors, we observe a correlation between predicted liver fat and imaging-based assessment of 0.58, again significantly lower than the correlation of >0.97 noted based on our machine learning model that included direct imaging data.

Manuscript Change(s): In response, we have updated the Methods and Results sections of the revised manuscript to more clearly describe this approach:

> *"As expected, the ability to quantify liver fat using direct imaging data was substantially higher than using clinical data alone.* ***For example, within the holdout testing dataset of 1,214 individuals, the correlation between body-mass index and liver fat was 0.42, improving to 0.58 in a model that incorporated 24 additional clinical factors and biomarker data (Figure S3)****."*

> *"To compare the performance of our machine learning, image-based model for liver fat quantification to an approach using clinical and anthropometric factors, we developed and tested a multivariable regression model...* ***Measurements at time of imaging assessment were available for BMI, height, weight, waist circumference, hip circumference, waist-to-hip ratio and systolic blood pressure and preferentially used in this regression analysis, while the remainder of predictors were measured at time of study enrollment."***

Additional details are provided in Figure S3 of the revised manuscript:

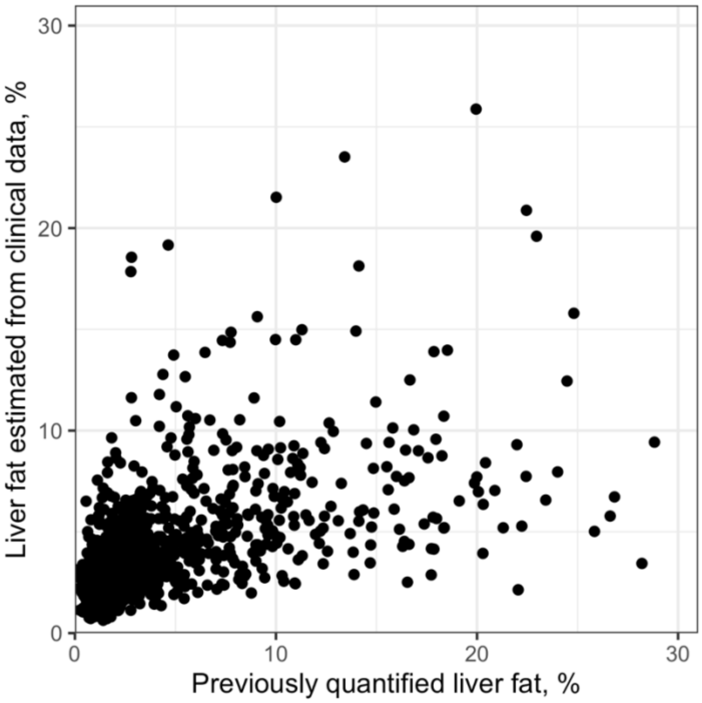

Figure S3. Prediction of liver fat using a variable dispersion beta regression model of clinical and anthropometric measurements, Related to STAR methods. In a held-out set of 1,214 participants of all self-reported ethnicities with previously-estimated liver fat, the Pearson correlation between the previously-quantified liver fat and liver fat estimated from a beta regression model using clinical and anthropometric measurements was 0.578 (95% CI 0.539-0.614; p-value=$3.8x10^{-109}$). Measurements that were at least nominally (p-value < 0.05) associated with liver fat in univariable analysis and therefore included in the beta regression model were: body-mass index, waist circumference, hip circumference, total body fat mass, total body fat percent, age at baseline, sex, height, weight, trunk fat mass, trunk fat percent, waist-to-hip ratio, LDL cholesterol, total cholesterol, HDL cholesterol, triglycerides, systolic blood pressure, alkaline phosphatase, alanine aminotransferase (ALT), aspartate aminotransferase (AST), ALT/AST, gamma glutamyltransferase, hemoglobin A1c, random glucose, and C-reactive protein. Lipid measures were adjusted for lipid-lowering medication use and blood pressure was adjusted for anti-hypertensive medication use, as previously described (Ehret et al., 2016; Patel et al., 2020).

Reviewer #2 Summary:
Author Reply: We appreciate and agree with the summary above.

Reviewer #2 Minor Comment 1:
Author Reply: We agree that this limitation warrants further emphasis.

Manuscript Change(s): In response, we have updated the opening sentence of the Discussion to read:

> *"Our analysis describing quantification of liver fat in 36,703 middle-aged participants in the UK Biobank* ***– the majority of whom were of European ancestry –*** *using a machine learning algorithm trained on a small subset with previously-quantified values has several implications for both biologic discovery and clinical medicine."*

We additionally assessed the predictive capacity of the machine learning model in hold-out test participants stratified by White versus non-White self-reported ethnicity noting comparable performance:

> *"Using a two-stage method with deep convolutional neural networks (see STAR Methods for details), we trained an algorithm to quantify liver fat that achieved near-perfect quantification: in hold-out testing datasets correlation coefficients were 0.97 and 0.99 and mean absolute errors were 0.50% and 0.41% in the two stages,* ***with comparable performance in self-reported White and non-White study participants (Figure S1).****"*

> *"Our results should be interpreted within the context of several potential limitations. First, participants of the UK Biobank imaging study tend to be healthier than the general population and* ***97%*** *were White based on self-reported ethnicity.* ***Although our algorithm for liver fat estimation appeared to perform comparably well in non-White participants (Figure S1)****, additional research is needed to investigate generalizability and transethnic portability."*

We have updated Figure S1 to show this additional assessment of the model in non-European ancestries

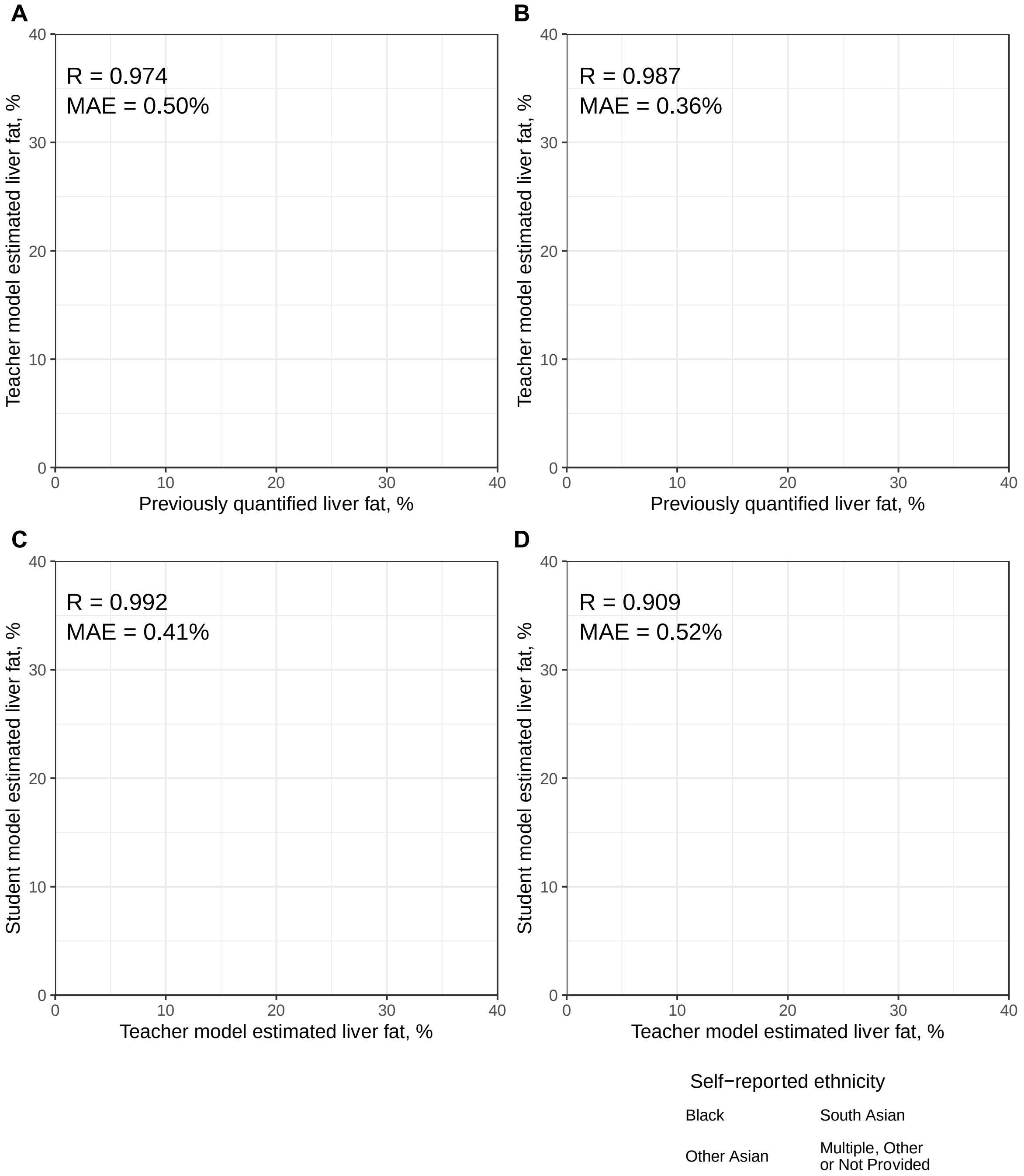

Figure S1. Comparison of previously-quantified liver fat with teacher-model inferences, and teacher-model inferences with student-model inferences, Related to STAR methods. A) In a held-out set of 1,214 participants with previously-quantified liver fat from gradient-echo imaging who were not used for model creation, the Pearson correlation between previously-quantified liver fat and liver fat inferred from the machine learning teacher model was 0.974 (95% CI 0.971-0.977; P<2.4x10^{-784}), and the mean absolute error was 0.50% (95%CI, 0.45-0.55%). B) The subset of testing individuals in A) with self-reported non-White ethnicity. C) In a separate held-out set of 383 samples with both gradient-echo and IDEAL imaging who were not used for model creation, the Pearson correlation between the teacher model inferred liver fat and the student model inferred liver fat was 0.992 (95% CI 0.990-0.993; P = 3.1 x 10^{-351}) and the mean absolute error was 0.41% (95%CI, 0.37-0.46%). D) The subset of testing individuals in C) with self-reported non-White ethnicity.

Reviewer #2 Minor Comment 2:

Author Reply: We agree that additional clarification is needed.

Manuscript Change(s): In response, we have edited the Results section of the revised manuscript to read:

> *"For the subset of 18,013 UK Biobank participants with both liver fat quantified and gene sequencing available, we next investigated whether rare inactivating DNA variants might affect liver fat or risk of steatosis.* ***Observed variants were included in this analysis based on minor allele frequency <0.1% and a prediction to cause premature truncation of a protein ('nonsense'), insertions or deletions that scramble protein translation ('frameshift'), or disruption of the messenger RNA splicing process ('splice-site') as annotated by the LOFTEE algorithm (Karczewski et al., 2020)."***

Reviewer #2 Minor Comment 3a:

Author Reply: We agree that additional clarification is needed. Principal components were developed centrally by the UK Biobank across all genotyped participants as previously described (Bycroft et al., 2018).

Manuscript Change(s): In response, we have updated the Methods section of the revised manuscript to read:

> *"Principal components of ancestry were calculated centrally by UK Biobank in all participants as previously described (Bycroft et al., 2018)."*

Reviewer #2 Minor Comment 3b:

Author Reply: We agree that standardization of all analyses to include ten principal components of ancestry is more appropriate.

Manuscript Change(s): In response, we have updated the polygenic score analyses to include the first ten principal components of ancestry, noting nearly identical results.

Reviewer #2 Minor Comment 4:

Author Reply: We agree that inclusion of sensitivity analysis based on both the U.S. and U.K. definitions for 'excessive' alcohol intake is warranted.

Manuscript Change(s): In response, we have added a sensitivity analysis excluding individuals exceeding the UK cut-off for excessive alcohol consumption (Table S6):

Table S6. Effects of eight liver fat CVAS variants on quantitative liver fat after adjusting for alcohol consumption, Related to Table 2.

					Unadjusted for alcohol consumption (original CVAS, n=32974)			Former alcohol consumers excluded, adjusted for number of weekly drinks (n=32062)			Former and excessive alcohol consumers (US guidelines) excluded (n=30216)			Former and excessive alcohol consumers (UK guidelines) excluded (n=23930)		
Lead Variant	Chr.	Position (hg19)	Nearest Gene	Effect Allele	Effect on liver fat (Beta)	SE	P-value	Effect on liver fat (Beta)	SE	P-value	Effect on liver fat (Beta)	SE	P-value	Effect on liver fat (Beta)	SE	P-value
Newly-Identified variants																
rs2642438	1	220970028	*MTARC1*	G	0.052	0.008	1.70E-09	0.050	0.009	8.70E-09	0.046	0.009	4.20E-07	0.052	0.010	3.20E-07
rs1229984	4	100239319	*ADH1B*	C	0.158	0.025	7.00E-10	0.153	0.026	3.40E-09	0.147	0.026	1.40E-08	0.116	0.029	5.30E-05
rs112875651	8	126506694	*TRIB1*	G	0.050	0.008	3.80E-10	0.053	0.008	3.70E-11	0.052	0.008	3.50E-10	0.057	0.009	8.30E-10
rs2250802	10	113921354	*GPAM*	G	0.054	0.009	1.40E-09	0.056	0.009	3.70E-10	0.056	0.009	2.10E-09	0.050	0.010	1.10E-06
rs56252442	19	18229208	*MAST3*	T	0.049	0.009	2.70E-08	0.050	0.009	1.70E-08	0.051	0.009	3.20E-08	0.039	0.011	1.60E-04
Previously-identified variants																
rs58542926	19	19379549	*TM6SF2*	T	0.289	0.015	2.80E-85	0.288	0.015	6.10E-83	0.283	0.015	1.60E-75	0.285	0.017	3.60E-61
rs429358	19	45411941	*APOE*	T	0.121	0.011	1.50E-29	0.120	0.011	3.90E-28	0.114	0.011	4.80E-24	0.102	0.013	3.90E-16
rs738409	22	44324727	*PNPLA3*	G	0.195	0.009	5.60E-95	0.194	0.010	1.10E-92	0.191	0.010	3.20E-84	0.175	0.011	6.10E-56

Chr., chromosome; SE, standard error; CVAS, common variant association study. Excessive alcohol intake, U.S. guidelines was defined as alcohol intake exceeding American Association for the Study of Liver Disease guidelines for NAFLD definition (Chalasani et al., 2018); excessive alcohol intake, U.K. guidelines was defined as alcohol intake exceeding the UK Chief Medical Officers recommendations (Department of Health, 2016).

We have also added this analysis to the Results:

> *"Given a known important role of alcohol intake on liver fat, we performed two sets of sensitivity analyses: first, we repeated the CVAS after exclusion of individuals who reported having stopped drinking alcohol or who reported alcohol consumption in excess of U.S. NAFLD* ***or U.K. guidelines****; second, we repeated the CVAS adjusting for self-reported number of alcoholic drinks consumed per week****. In both cases, results for the 8 variants identified were largely similar, suggesting that these variants have a consistent impact on liver fat independent of alcohol consumption (Table S6). For the p.H48R missense variant in ADH1B, the effect size was somewhat reduced but an association with increased liver fat remained in all sensitivity analyses (p-values 5.3x10^{-5} to 3.4x10^{-9}).****"*

Reviewer #2 Minor Comment 5:
Author Reply: We agree that exclusion of individuals with known viral hepatitis B or C infection is appropriate, understanding that such infections were documented in fewer than 0.25% of UK Biobank participants.

Manuscript Change(s): In response, we have updated the polygenic score analysis to exclude an additional 244 participants of the 362,096 previously studied (0.1%), noting nearly identical results.

This additional exclusion is described in the Methods section of the revised manuscript:

> *"We tested for association between the score and incident disease occurrence after UK Biobank enrollment using a Cox model in the same set of individuals used to test associations between single CVAS variants and NAFLD/NASH. We excluded individuals who had any of the four diseases investigated* ***or hepatitis B or C infection*** *documented at time of enrollment, resulting in 361,852 participants in the analysis."*

Reviewer #2 Minor Comment 6:
Author Reply: We agree that this limitation warrants further clarification. As with other circulating biomarkers, ALT and AST concentrations demonstrate moderate fluctuations over time (10-30% day-to-day for ALT, Kim et al., 2008). Consistent with this, we examined ALT and AST in the subset of ~16,500 UK Biobank participants had biomarkers assessed at both study enrollment and a second visit ~4 years later. Among these individuals median percent changes for ALT and AST were 21% and 13% respectively. We also note that associations of the CVAS variants with ALT and AST were measured in an independent group of UK Biobank participants that were not part of the imaging sub-study.

Manuscript change(s): In response, we have clarified this issue in the Results and Methods sections of the revised manuscript:

> *"We also examined the association of liver fat with circulating biomarkers* ***collected at time of enrollment****, noting that circulating triglycerides, liver-associated aminotransferases and glycemic indices were all significantly increased in those with steatosis."*

> *"Beyond association with liver fat indices, we sought additional validation of the variants identified by CVAS using liver biomarkers* ***assessed at time of study enrollment****"*
>
> *"As part of the study protocol, a subset of individuals underwent detailed imaging including abdominal MRI (Littlejohns et al., 2020) between 2014 and 2019,* ***an average of 9.3 years after enrollment visit.****"*

Reviewer #2 Minor Comment 7:
Author Reply: We agree that additional clarification is needed. As described in the Methods section, UK Biobank participants were imaged using two distinct study protocols – a gradient echo protocol that included 10 images and the IDEAL protocol that included 36 images.

Manuscript change(s): In response, we have edited the Methods section of the revised manuscript to read:

> *"The gradient echo protocol consisted of acquiring 10 images (Wilman et al., 2017); to avoid potential errors in estimation that could arise from using a different number of images, we restricted the participants used for model training to individuals who had 10 images, resulting in 3,210 used for model training and 1,215 held out for model testing…*
>
> *To estimate liver fat in participants imaged using the IDEAL protocol, we also trained a 2D CNN "student" model in the participants who had undergone both the gradient echo and IDEAL imaging protocols.* ***The IDEAL protocol included 36 images*** *with largest image pixel value < 1024; of the 1,441 individuals who had both imaging protocols and these 36 images, 1,057 were used for training and 384 were held out for testing."*

Reviewer #2 Minor Comment 8:
Author Reply: We agree that this information is of interest. Within routine clinical practice, liver fat is rarely quantified using liver biopsy or the (imaging-based) gold-standard of proton density as assessed by MRI. We agree that deployment of machine learning algorithms to quickly and accurately quantify liver fat for scans obtained within clinical practice may be of significant value.

Manuscript change(s): In response, we have edited the Discussion section of the revised manuscript to read:

> *"Previous efforts have similarly shown feasibility of using a convolutional neural net framework to automate liver fat quantification using CT or MRI images in clinical practice (Wang et al., 2019). Such efforts may be of particular value for liver fat, since within routine clinical practice, liver fat noted from ultrasound or CT imaging is typically reported in qualitative rather than quantitative terms that lack precision and accuracy (Zhang et al., 2018)."*

Reviewer #3 Summary:
Author Reply: We appreciate and agree with the summary above.

Reviewer #3 Major Comment 1:
Author Reply: We agree that inclusion of mean absolute error as a second measure of model performance is appropriate. For the two models, we observed a mean absolute error of 0.50% (95%CI, 0.45-0.55%) for the teacher model and 0.41% (95%CI, 0.37-0.46%) for the student model.

Manuscript Change(s): In response, we have incorporated evaluation of the ML model using mean absolute error into the Results section:

> *"Using a two-stage method with deep convolutional neural networks (see STAR Methods for details), we trained an algorithm to quantify liver fat that achieved near-perfect quantification: in hold-out testing datasets correlation coefficients were 0.97 and 0.99* ***and mean absolute errors were 0.50% and 0.41% in the two stages****, with comparable performance in self-reported White and non-White study participants (Figure S1)."*

Data on mean absolute error is additionally included in the legend of Figure S1 of the revised manuscript:

> *"Figure S1. Comparison of previously-quantified liver fat with teacher-model inferences, and teacher-model inferences with student-model inferences, Related to STAR methods. A) In a held-out set of 1,214 participants with previously-quantified liver fat from gradient-echo imaging who were not used for model creation, the Pearson correlation between previously-quantified liver fat and liver fat inferred from the machine learning teacher model was 0.974 (95% CI 0.971-0.977; $P<2.4x10^{-784}$), and* ***the mean absolute error was 0.50% (95%CI, 0.45-0.55%)****… C) In a separate held-out set of 383 samples with both gradient-echo and IDEAL imaging who were not used for model creation, the Pearson correlation between the teacher model inferred liver fat and the student model inferred liver fat was 0.992 (95% CI 0.990-0.993; $P = 3.1 \times 10^{-351}$) and* ***the mean absolute error was 0.41% (95%CI, 0.37-0.46%).****"*

We have additionally provided description of this approach in the Methods section of the revised manuscript:

> *"Performance on the held-out testing sets was assessed based on Pearson correlation coefficient* ***and mean absolute error*** *for each model (Figure S1)."*

Reviewer #3 Minor Comment 2:
Author Reply: We agree that the precise p-values do not add to the study.

Manuscript change(s): In response, we have updated Table 1 and comparisons of clinical characteristics, replacing p-values < 0.001 as '<0.001' as suggested.

Reviewer #4 Summary:
Author Reply: We appreciate and agree with the summary above.

Reviewer #4 Comment 1:
Author Reply: We agree cross-validation within the training dataset is an additional approach to confirming lack of overfitting and model performance.

As outlined below, we have updated the revised manuscript to include 10-fold cross validation for both the teacher and student models, confirming performance nearly identical to the models that included the full training dataset:

- Teacher model
    - Full training dataset:
        - correlation coefficient of 0.974
        - MAE of 0.50%
    - Cross-validation (mean across 10-folds):
        - correlation coefficient of 0.975
        - mean absolute error 0.50%
- Student model
    - Full training dataset:
        - correlation coefficient of 0.992
        - MAE of 0.41%
    - Cross-validation (mean across 10-folds):
        - correlation coefficient of 0.983
        - mean absolute error 0.58%

Performances in each of the 10 folds are indicated in the manuscript text below.

Manuscript change(s): In response, we have included this information in the Methods section of the revised manuscript:

> *"As an additional sensitivity analysis, we performed 10-fold cross validation within subsets of the training datasets, noting nearly identical performance as for the model developed using the full training datasets. For the teacher model, we observed a mean Pearson correlation coefficient across each of 10 folds of 0.975 (values in each fold: 0.970, 0.976, 0.976, 0.976, 0.976, 0.977, 0.976, 0.976, 0.974, 0.976) and an average mean absolute error across each of 10 folds of 0.50% (values in each fold: 0.57%, 0.49%, 0.53%, 0.46%, 0.50%, 0.49%, 0.48%, 0.50%, 0.52%, 0.51%). For the student model, we observed a mean Pearson correlation coefficient across each of 10 folds of 0.983 (values in each fold: 0.985, 0.985, 0.978, 0.974, 0.984, 0.982, 0.981, 0.986, 0.985, 0.987) and an average mean absolute error of 0.58% (values in each fold: 0.53%, 0.54%, 0.69%, 0.65%, 0.56%, 0.58%, 0.62%, 0.56%, 0.52%, 0.52%)."*

Reviewer #4 Comment 2:
Author Reply: We agree that potential imperfect assessment of liver fat in our study may impact downstream association results, but we emphasize that this would likely bias results toward the null rather than introduce false-positive associations.

Although the CNN models do not produce individual-level confidence assessments – and we observed excellent accuracy in all participant subgroups of a hold-out testing dataset studied – we set out to confirm that a small portion of samples with highly inaccurate values did not drive our genetic association study results.

For each of 8 common variants associated with liver fat, we performed 10-fold cross validation by systematically excluding 10% of the data within each fold. As noted in the table below, effect estimates were highly consistent and within a narrow range for all such variants:

		CVAS effect on liver fat: Beta (95% CI) - as reported in Table 2	Cross validation of effect on liver fat: Min-Max range of Betas across 10 folds
Newly-identified variants			
rs2642438	*MTARC1*	0.052 (0.035-0.068)	0.048-0.055
rs1229984	*ADH1B*	0.158 (0.108-0.208)	0.138-0.166
rs112875651	*TRIB1*	0.050 (0.034-0.066)	0.045-0.053
rs2250802	*GPAM*	0.054 (0.037-0.071)	0.049-0.060
rs56252442	*MAST3*	0.049 (0.032-0.067)	0.044-0.049
Previously-identified variants			
rs58542926	*TM6SF2*	0.289 (0.260-0.318)	0.277-0.298
rs429358	*APOE*	0.121 (0.100-0.143)	0.113-0.124
rs738409	*PNPLA3*	0.195 (0.176-0.213)	0.186-0.202

References

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L.T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., et al. (2018). The UK Biobank resource with deep phenotyping and genomic data. Nature 562, 203–209.

Chalasani, N., Younossi, Z., Lavine, J.E., Charlton, M., Cusi, K., Rinella, M., Harrison, S.A., Brunt, E.M., and Sanyal, A.J. (2018). The diagnosis and management of nonalcoholic fatty liver disease: Practice guidance from the American Association for the Study of Liver Diseases. Hepatol. Baltim. Md 67, 328–357.

Department of Health (2016). UK Chief Medical Officers' Low Risk Drinking Guidelines (London).

Ehret, G.B., Ferreira, T., Chasman, D.I., Jackson, A.U., Schmidt, E.M., Johnson, T., Thorleifsson, G., Luan, J., Donnelly, L.A., Kanoni, S., et al. (2016). The genetics of blood pressure regulation and its target organs from association studies in 342,415 individuals. Nat. Genet. 48, 1171–1184.

Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alföldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P., et al. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. Nature 581, 434–443.

Kim, W.R., Flamm, S.L., Bisceglie, A.M.D., and Bodenheimer, H.C. (2008). Serum activity of alanine aminotransferase (ALT) as an indicator of health and disease. Hepatology 47, 1363–1370.

Littlejohns, T.J., Holliday, J., Gibson, L.M., Garratt, S., Oesingmann, N., Alfaro-Almagro, F., Bell, J.D., Boultwood, C., Collins, R., Conroy, M.C., et al. (2020). The UK Biobank imaging enhancement of 100,000 participants: rationale, data collection, management and future directions. Nat. Commun. 11, 2624.

NHLBI Expert Panel (1998). Clinical Guidelines on the Identification, Evaluation, and Treatment of Overweight and Obesity in Adults--The Evidence Report. Obes. Res. 6 Suppl 2, 51S-209S.

Patel, A.P., Wang, M., Fahed, A.C., Mason-Suares, H., Brockman, D., Pelletier, R., Amr, S., Machini, K., Hawley, M., Witkowski, L., et al. (2020). Association of Rare Pathogenic DNA Variants for Familial Hypercholesterolemia, Hereditary Breast and Ovarian Cancer Syndrome, and Lynch Syndrome With Disease Risk in Adults According to Family History. JAMA Netw. Open 3, e203959.

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., et al. (2015). UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. PLOS Med. 12, e1001779-10.

Wilman, H.R., Kelly, M., Garratt, S., Matthews, P.M., Milanesi, M., Herlihy, A., Gyngell, M., Neubauer, S., Bell, J.D., Banerjee, R., et al. (2017). Characterisation of liver fat in the UK Biobank cohort. PloS One 12, e0172921-14.

Zhang, Y.N., Fowler, K.J., Hamilton, G., Cui, J.Y., Sy, E.Z., Balanay, M., Hooker, J.C., Szeverenyi, N., and Sirlin, C.B. (2018). Liver fat imaging—a clinical overview of ultrasound, CT, and MR imaging. Br. J. Radiol. 91.

Referee reports, second round of review

Reviewer #1: This is an interesting study, that describes a new algorithm for measuring liver fat from MRI images. This is likely to be very useful for research. The authors validate the algorithm convincingly - using training data from 4000 individuals with alternative imaging based measures - and use it to increase the power for genetic studies. The study has been through review already and i think is interesting so I do not feel strongly about the below minor comments that may or may not help the reader. I have only one major comment:
1. I found myself very convinced of the validity of the measures, but was not sure of their utility, in that how much better do they do at predicting future liver disease, T2D, CVD etc compared to much simpler and cheap measures such as LFTs ? The MRI measures are clearly better at measuring liver fat than BMI and biomarkers, but how does this translate to outcomes ? This seemed to be missing as the paper does not compare the prediction of future NAFLD to baseline LFTs I don't think, but goes straight on to the GWAS. If the paper is simple about a more accurate measure for research it is fine as is, but the authors talk about utility to clincians so they might like to consider this point.
Minor comments:
2. Summary suggests 8 variants in the GRS can "strongly predict" future liver disease but an odds ratio of 1.33 per SD or AUC going from 0.55 to 0.6 is not strongly predictive - strongly associated perhaps ?
3. Lines 125-158 talk about BMI as a poor predictor of liver fat compared to direct imaging. But what about ALT and other LFTs ? the comparisons to make this point should be with the simple things that can be measured but LFTs fall under that category.
4. The "teacher-student" terminology is a little unusual ? most people in the field i believe will be more familiar with "training-test" datasets ?
5. Line 91 - you may wish to clarify what you mean by "unascertained" - i think you mean "unascertained for any clinical indication" or "population ascertained"
Tim Frayling

Reviewer #2: All of my concerns have been addressed and the paper is stronger.

Reviewer #3: The authors have satisfactorily addressed my comments from the previous round of reviews.

Reviewer #4: Comments enter in this field will be shared with the author; your identity will remain anonymous.

The authors have addressed my concerns. I do not have additional questions. Therefore, I recommend acceptance.

Author response to the second round of review

Reviewer #1 Summary:
Author Reply: We appreciate and agree with the summary above.

Reviewer #1 Comment 1:
Author Reply: We agree that these questions are of interest. In our paper, we demonstrate that imaging based assessment identifies 17% of participants with liver steatosis. Only a small fraction of these individuals carried a diagnosis code of this condition in the medical record, and measures such as BMI, LFTs, or other clinical factors could not be used to reliably identify them. As such, imaging-based approaches – especially if data is available as latent information in the medical record – to identify those with steatosis may prove useful in enhancing diagnosis rates.

A separate (and also important) question is whether imaging-based assessment of liver fat outperforms clinical markers in predicting risk of *future* disease. Here, the UK Biobank dataset has important limitations for two reasons:

First, LFT measurement and abdominal MRI did not occur at the same time and occurred in different subsets of participants, and it is thus not possible to do the appropriate ‘apples to apples’ comparison:

- ALT last measured at ‘Visit 2’, 2012-2013; n = 17,863
- Liver fat measured by MRI at ‘Visit 3’, 2014-2019; n = 36,703

Second, because the imaging occurred only recently, participants have been followed for a median of only 2.9 years, limiting the number of incident events for analysis:

- Liver-related incident disease outcomes in imaged population
    - Nonalcoholic steatohepatitis: N = 13
    - Cirrhosis: N = 24
    - Hepatocellular carcinoma: N = 9
- Non-liver related incident disease outcomes in imaged population
    - Type 2 diabetes N = 179
    - Coronary artery disease N = 343

Despite these limitations, we conducted exploratory analyses that compared the hazard ratios for incident disease according to ALT concentrations and MRI-measured liver fat:

Disease	N events/N individuals	Predictor	Median follow-up years	HR per SD predictor*	P-value	HR for high liver fat or high ALT**	P-value
Nonalcoholic Steatohepatitis	13/36686	Liver fat	2.9	1.55 (1.21-1.99)	5.6E-04	6.82 (2.20-21.12)	8.7E-04
	10/17861	ALT	8.1	1.35 (1.16-1.57)	1.1E-04	6.47 (1.81-23.06)	0.004
Cirrhosis	24/36656	Liver fat	2.9	1.29 (0.99-1.68)	0.062	2.10 (0.89-4.96)	0.09
	47/17844	ALT	8.1	1.32 (1.22-1.44)	1.1E-11	4.63 (2.60-8.24)	2.0E-07
Hepatocellular Carcinoma	9/36697	Liver fat	2.9	1.16 (0.66-2.04)	0.61	2.41 (0.59-9.86)	0.219
	9/17861	ALT	8.1	1.32 (1.10-1.59)	0.003	6.52 (1.72-24.77)	0.006
Coronary Artery Disease	343/35398	Liver fat	2.9	1.12 (1.02-1.22)	0.016	1.32 (1.02-1.69)	0.032
	467/17105	ALT	8.1	1.00 (0.91-1.10)	0.955	1.02 (0.80-1.31)	0.876
Diabetes	179/34739	Liver fat	2.9	1.63 (1.53-1.74)	4.3E-49	5.11 (3.80-6.88)	5.9E-27
	314/16720	ALT	8.1	1.23 (1.17-1.28)	1.7E-19	2.72 (2.15-3.45)	1.2E-16

Analyses adjusted for sex, age at imaging or repeat visit, age at imaging or repeat visit squared, birth year, and (liver fat only) MRI machine serial number. *Standard deviation liver fat = 4.3%, standard deviation ALT = 12 U/L. **High liver fat, liver fat > 5.5%; high ALT, ALT > 33 U/L if male; >25 U/L if female (Kwo et al., 2017).

These results suggest that liver fat and ALT provide both overlapping and distinct information for disease risk. The stronger predictive power of liver fat for coronary artery disease and type 2 diabetes is consistent with metabolic dysfunction such as insulin resistance predisposing to both increased liver fat and these diseases. Similarly, the stronger predictive power of ALT for cirrhosis and hepatocellular carcinoma is consistent with elevated ALT resulting from increased liver fat as well as additional drivers of liver disease.

However, additional research is needed in cohorts where ALT and liver fat are measured simultaneously in the same individuals, with sufficient follow-up time, to enable a rigorous head-to-head comparison.

Manuscript Change(s): In response, we have edited the Discussion section of the revised manuscript to read:

> *"Our results should be interpreted within the context of several potential limitations...Third, because imaging of UK Biobank participants occurred recently and not at time of enrollment,* ***we were not able to directly compare the predictive power of liver fat versus other clinical or biomarker predictors with respect to future risk of cardiometabolic or liver diseases****."*

Reviewer #1 Minor Comment 2:
Author Reply: We agree that this description is more appropriate.

Manuscript Change(s): In response, we have updated the Summary to read:

> *"A common DNA variant association study of liver fat confirmed three known associations and newly identified variants in or near the MTARC1, ADH1B, TRIB1, GPAM and MAST3 genes. A polygenic score integrating these eight variants* ***was strongly associated with future risk*** *of chronic liver diseases."*

Reviewer #1 Minor Comment 3:
Author Reply: We agree that additional clarification is warranted. Liver function tests – including ALT, AST, GGT and ALT/AST which were measured at the baseline visit (an average of 9.3 years prior to the imaging visit) – were included as 4 of the 24 clinical parameters in addition to BMI used to quantity liver fat from clinical data alone (lines 118-121 and Figure S3), resulting a Pearson correlation of only 0.58 with previously-quantified liver fat.

An exploratory analysis that explores each of these factors in isolation similarly notes modest correlation between these biomarkers and liver fat assessed on direct imaging:

- Pearson correlation values between LFT concentrations and liver fat:

ALT	r = 0.25
AST	r = 0.12
GGT	r = 0.18
ALT/AST	r = 0.35

For lines 125-158, we focused on BMI and WHR as these were remeasured in participants on the day of the imaging visit.

Manuscript Change(s): In response, we have added additional details about the model incorporating clinical parameters:

> *"As expected, the ability to quantify liver fat using direct imaging data was substantially higher than clinical data alone. For example, within the holdout testing dataset of 1,214 individuals, the correlation between body-mass index and liver fat was 0.42, improving to 0.58 in a model that incorporated 24 additional clinical factors and biomarker data* ***including liver-related biomarkers such as alanine aminotransferase*** *(Figure S3)."*

Reviewer #1 Minor Comment 4:
Author Reply: We agree that the 'teacher-student' semantics may not be as familiar to the *Cell Genomics* audience and have sought to clarify two distinct concepts for the readership:

- 'Teacher-student' model – a commonly used approach in machine learning that allowed development of algorithms that work across two different imaging protocols used for UK Biobank participants
- 'Training-test' datasets – the approach used to train a machine learning model and test it in an independent holdout dataset, with accuracy confirmed for each of two imaging protocols

Reviewer #1 Comment 5:
Author Reply: We agree that additional clarification here is appropriate.

Manuscript Change(s): In response, we have updated the introduction:

> *"Second, the association of clinical risk factors with hepatic steatosis, as well as the ability to predict liver fat content without direct imaging, have not been fully characterized in large studies of individuals* ***not ascertained for any specific clinical indication.****"*

References:

Kwo, P.Y., Cohen, S.M., and Lim, J.K. (2017). ACG Clinical Guideline: Evaluation of Abnormal Liver Chemistries. Am. J. Gastroenterol. 112, 18–35.