Peer Review File

Article Information: https://dx.doi.org/10.21037/jgo-22-882

REVIEWER A

Comment 1: "The introduction is too long and much of this belongs in the discussion. The intro needs to be shortened to 1- 1.5 pages and the detailed analysis of prior studies should be saved for the discussion and how it relates to the findings of this study. This manuscript reads as more as a review paper instead of setting up the topic to be investigated."
Reply 1: We agree with the request and have shortened the introduction to to 1-1.5 pages. We have also moved sections from the introduction into the discussion section.
Changes in the text: *Introduction***:** Page 3-4, lines 54-55 and 67-83**;** *Discussion***:** page 11, lines 251-261.

Comment 2: "In the table why are CP class B and C lumped together? I would separate these."
Reply 2: Thank you for this suggestion. We have separated CP Class B and C.
Changes in the text: Added to Table 1.

Comment 3: "In table 1 for other treatments, when were these given? Before Y90, concurrently, or after?"
Reply 3: Regarding the timing of initiation of other HCC treatments, they could have been given before or after Y90. Given the nature of HCC and the heterogeneity of each patient's case regarding lesion(s) (size, location), liver function, etc, treatment decisions and the modality of treatment was based on the decision of VCU's multidisciplinary liver tumor board. Some of our patients were on the transplant list, and they received Y 90 or other locoregional therapies as a bridge to transplant. We agree this is a limitation to our study but again highlights the real- world experience.
Changes in the text: Page 17, lines 425-427.

Comment 4: "One of my main concerns with the study is the following: "Patients were included in the study if they were at least 18 years old and 106 had a diagnosis of primary HCC that was treated with Y90 at VCU Health between 2008 and 107 2015." This encompasses an extremely broad and heterogenous group of patients, treating physicians (potentially), treatment parameters, and types of HCC. The efficacy and long term outcomes of Y90 can vary significantly based on the type of HCC being treated. For example, large greater than 10cm HCCs with vascular invasion are treated much different from small 2cm HCC, which are treated different from multifocal HCC isolated to one lobe, which is treated different from multifocal HCC in both lobes of the liver. Also, patients being bridged to liver transplant are treated different from those undergoing definitive therapy or those that are not transplant candidates. Y90 is now a mature treatment technique that is tailored to very specific to tumor characteristics, patient factors, etc, and this can result in very different outcomes. Furthermore, Y90 itself is versatile in that it can be used as glass/resin and one can

do ablative dosing versus palliative dosing depending on the clinical context. With so much heterogeneity in the possible tumors treated and possible ways to utilize Y90 it makes it very challenging to interpret the rest of the results. The authors need to provide a much more granular look at the tumors treated as well as the Y90 parameters used in order to put these results into context."

Reply 4:

We agree with the Reviewer's comments regarding the versatility of the Y90 and how it can be used as modality in various contexts of HCC, ranging from a bridge-to-transplant to palliative care measures.

We recognize the heterogeneity of our data in the use of Y90 at our institution.
In treatment approaches with Y90 at our institution, if there is a single small lesion, our IR team will often do a super selective embolization with a tumor dose exceeding 250 or 300 gray if possible. This would be akin to a definitive or ablative or curative treatment. Our IR team often pursues this when possible, whether they are transplant candidates or not, as it is not always clear if the patient will get to transplant.

We agree there are several factors that go into the decision of dosing, including the patient's anatomy, number of lesions, size of lesions, percentage of shunt to the lungs, underlying liver function, goal of treatment. And even within our IR group, there is variability in treating with some radiologists more custom in their dosing and some more routine. Therefore, we agree it is difficult to make conclusions in a modern treatment scheme due to variability in treatment approaches.

Because our study is reflective of real- world analysis and real- word experience, and not a controlled trial, we will not be able to provide these granular details, but have added text to acknowledge this as a limitation to our study.

Changes in the text: Pages 17, lines 407-409, Page 17, lines 425-427.

Comment 5: "OS and PFS from date of Y90 treatment was 15 m and 3 m. What is the explanation for such a low PFS after Y90 treatment?"

Reply 5: We acknowledge that the PFS from date of Y90 was unusually shorter compared to other studies. We suspect that the difference in PFS may be related to our radiologists using the Liver imaging Reporting and Data System (LiRADS) in interpreting HCC treatment response instead of the RECIST criteria. Because our study is focused on real-world practices, by using LiRADS we were focused on actual clinical decision making rather than traditional stable disease criteria used in clinical trials, which can inflate durability of benefit.

We suspect that Y90 may be associated with shorter PFS because of this. This makes our study a potential strength.

VCU is a very active liver transplant center and our radiologists utilize the LiRADS criteria for diagnosis of HCC as well as response to treatment, which is standard of care and practice in the United States. In 2011, the American College of Radiology launched the use of LiRADS to standardize interpretation and reporting of diagnosis of HCC lesions as well as to standardize the assessment of response after treatment with locoregional therapy.

We acknowledge that mRECIST is used as surrogate endpoints in clinical trials and its criteria apply to the concept of viable tumors, which corresponds to the portion of tumors that shows enhancement after intravenous contrast injection. All liver transplant centers in the United States use LI-RADS. The LI-RADS treatment response (LR-TR) algorithm extends the definition of viable tumor by adding new imaging features such as washout appearance and enhancement, similar to pretreatment. Whereas, mRECIST considers arterial phase hyperenhancement (APHE) as the only characteristic of a viable tumor. Moreover, in the LR-TR algorithm, the treated observation can be classified as an equivocal category when the distinction between viable tumor and the expected posttreatment enhancement is uncertain. Therefore, the LR-TR algorithm uses a ternary system that categorizes the treated observations as LR-TR viable (probably or definitely viable), equivocal (equivocally viable), or nonviable category. Similar to mRECIST, the LI-RADS algorithm is based on unidimensional measurements of the largest enhancing component of a treated tumor, excluding areas of nonenhancement, an approach that shows high reproducibility in response categories and better prognostication when compared with traditional RECIST version 1.1, after local-regional therapy. The LI-RADS algorithm expands on the mRECIST approach not only by defining viable disease but also by providing nonevaluable, equivocal, and nonviable treatment response categories. Unlike mRECIST, the LI-RADS treatment response categories are assigned on a lesion-by-lesion basis and are not assigned to the whole liver or patient.

With these differences between the LR-TR algorithm and mRECIST criteria, there have been studies that have compared the diagnostic performance between the two criteria with reported results varying considerably. One recent meta- analysis comparing mRECIST and LiRADS criteria suggested that LiRADS criteria have better specificity than mRECIST, without a significant difference in sensitivity for the diagnosis of pathologically viable HCC after LRT. Thus, these differences could explain why the PFS in this study is lower.

(Kim DH, Kim B, Choi JI, Oh SN, Rha SE. LI-RADS Treatment Response versus Modified RECIST for Diagnosing Viable Hepatocellular Carcinoma after Locoregional Therapy: A Systematic Review and Meta-Analysis of Comparative Studies. Taehan Yongsang Uihakhoe Chi. 2022 Mar;83(2):331-343. doi: 10.3348/jksr.2021.0173. PMID: 36237934).

We have added this explanation to the manuscript.
Changes in the text: Pages 11-12, lines 266-290.

Comment 6: "Analysis 169 of both OS and PFS based on cancer stage also showed no statistical significance. What is the explanation for this finding? It seems peculiar to me that cancer stage had no impact on OS and PFS? Also, with regard to cancer stage, where are these results? I do not see a tabulation of the cancer stages treated. This is critical to know."
Reply 6: We appreciate the the reviewer's comments and agree that our cohort comprises of various stages of HCC and that this could have impact on OS and PFS. To provide further granularity, we have listed in Table 1 the % of patients with their AJCC TNM.

We also have provided additional analysis re: OS/ PFS and cancer staging:

1) We have included here an additional table and figure on OS by cancer stage (Table 2a/ Figure 1a). This did not show differences in OS by cancer stage. This being an observational study, the explanation for this could be due to our small sample size.

2) We have also included a table/ figure on PFS by cancer stage (Table 2b/ Figure 1b). We do demonstrate here that PFS differs by cancer stage between Stage 1 and 3. Stage 1 had greater median PFS than Stage 3. No other groups differences reached statistical significance.

3) We also had already performed a multivariate analysis showing correlation between older age, high cancer stage, higher MELD and CP class and presence of PVT with significantly lower OS. Those with advanced stage disease at diagnosis and higher CP score also had significantly shorter PFS. This was already stated in the manuscript on page 9, lines 197-202.

Changes in the text: Page 2, lines 34-36; Pages 7-8, lines 165-166; Page 10, lines 239-248.

Comment 7: "Multiple comparisons are made for a median OS of 12 months. How was the benchmark of 12 months selected? (lines 179-186)."
Reply 7: We selected 12 months arbitrarily as a reasonable marker of time to assess benefits. As we know patients with liver cancer and Childs B/C have terrible prognosis and much shorter survival time, we chose 12 months as this time period would likely be able to capture any benefit.

Comment 8: "The information in 194-198 means very little without understanding how Y90 was performed and for what purpose. These parameters will vary significantly if we are treating 1cm HCC versus 10cm HCC with portal invasion vs bilobar treatment."
Reply 8: We appreciate the Reviewer's feedback regarding how Y90 was performed and for what purpose. We have included in Table 1 the number of patients who received single lobe vs. bilobar treatment. Majority 89% received single lobe treatment. We have also included data in Table 1 regarding number of patients with portal invasion which involved 15% of our patients.

Unfortunately, given the significant heterogeneity of the tumor size in our patient population, we were not able to capture the exact details of tumor size. This has been commented as a limitation in our study.
However, details of cancer staging has been placed in Table 1 as an addition.

Changes in the text: Page 17, line 425-427.

Comment 9: "This may 228 reflect the utilization of RECIST in clinical trials in contrast to clinical radiology practice, as was. What do the authors mean by "clinical radiology practice" to determine progression?"
Reply 9: We appreciate the author's comments regarding the definition of "clinical radiology practice." Please see our response in question 5 that further clarifies the use LiRADS vs. RECIST.

Changes in the text: Page 11-12, lines 266-290.

Comment 10: "The increasing MELD and CP scores over time likely reflects a combination of REILD and progression of HCC, 269 both of which can worsen synthetic liver function. The authors stated that only 9 patients had REIL so why would it have such a big influence on MELD and CP scores for so many other patients?"
Reply 10: We thank our reviewer for this very important question. We agree that only 9 patients were formally diagnosed with REILD. We failed to mention that other patients did develop liver decompensation that did not fit criteria for REILD. This includes categories under Table 8 of "worsening liver failure, including 9% ascites/fluid retention, 4% hepatic encephalopathy. This has been added to the text.

Changes in the text: Page 2, line 42**;** Page 4, lines 87-88; Page14, 336-338.

Comment 11: "The section on future directions in the era of immunotherapy should not be a part of this manuscript. It deters from the overall point and beings to discuss an entirely new topic. This should be minimized to 1-2 sentences if the authors wish to discuss it."
Reply 11: We appreciate the reviewer's feedback regarding this section. We have deleted this section.
Changes in the text: Page 15-16, lines 377-406.

Comment 12: "Rather, progression was based on standard readings in clinical radiology practice. This is a very significant limitation to this paper and one that I think needs to be revised. What does standard readings in clinical radiology practice mean? If we use these

readings how can we truly compare these results to the rest of the literature. There is often a disconnect between Y90 treatment efficacy and the reading of the diagnostic radiologist. In this case the results are skewed in the negative by the reading radiologist. The reading radiologist does not have enough time to provide a detailed report and that is why we retrospectively analyze data for scientific purposes. Furthermore, the reading radiologist will vary significantly from institution to institution."

Reply 12: We acknowledge the Reviewer's comments regarding the importance of defining what standard readings in clinical radiology practice is.

This is answered also in **comment/ reply 4 and 5** regarding standard practice with the use of LiRADS. This again reflects the differences of evaluating HCC treatment response and progression between clinical trial parameters (RECIST) and real world practice/ standard of care with LiRADS.

The purpose of LiRADS is to prevent interobserver variation between radiologists and between institutions. LiRADS standardizes interpretation and reporting of diagnosis of HCC lesions as well as assessing the response to HCC lesions after treatment with locoregional therapy.

We acknowledge the challenges our paper may be pose when comparing it to prior studies using RECIST criteria. However, we feel our paper represents real world practice parameters and is a stepping stone adding a new perspective and additional insight to HCC treatment response that is applicable to day-to-day clinical practice in liver transplant center. By utilizing LiRADS criteria we are applying standardized and nationally accepted methodology in the US in a real-world setting.

Changes in the text:

Comment 13: There is no mention throughout this manuscript of Y90 treatment parameters which is a major limitation.

Reply 13: We appreciate the Reviewer's comments regarding our lack of Y 90 treatment parameters.

The versatility of the Y90 and how it can be used as modality in various contexts of HCC ranging from a bridge-to-transplant to palliative care measures. Due to the heterogeneity of our data in the use of Y90 at our institution, we will not be able to provide these granular details, but have added text to acknowledge this as a limitation to our study.

Please see our **comment/ reply 4** regarding our general approach to Y 90 treatment.

Changes in the text: Pages 17, lines 425- 427.

Comment 14: Overall, I believe this manuscript needs significant fine-tuning. I understand that the authors are trying to paint a picture of "real world Y90", but in doing so they make

the results very challenging to interpret. Y90 is a mature locoregional HCC treatment modality that is tailored to patients based on their tumor characteristics and patient factors (labs, ECOG, treatment intent). We cannot simply lump together resin Y90, glass Y90, bilobar HCC, solitary HCC, multifocal HCC, PVT, segmentectomy treatment, lobar treatment etc. With all of these lumped together it makes the results extremely difficult to interpret and it makes them very challenging to compare with the rest of the literature. Furthermore, it doesn't really help the reading physician understand how these results relate to their clinical practice or how to use them to improve their practice. Furthermore, the authors us the AJCC TNM classification which is not typically used in descriptive Y90 studies. They need to also include RECIST scoring system and use that in follow-up as well so that the results can be compared to what is out there already.

Reply 14: We appreciate the reviewer's important feedback and acknowledge the challenges that are seen in our manuscript.

We have added to Table 1 additional descriptive and more granular data including cancer staging, tumor location (unilobar/ bilobar/ diffuse), PVT (bland vs. malignant)/ (right/ left/ main branch), treatment with Y90 with single or bilobar. We also acknowledge the AJCC TNM classification is not used in clinical trials when studying locoregional therapy such as Y90; therefore, it may be difficult to compare with other studies. However, our study may provide further insight and demonstrate the differences between real world practice and clinical trials.
Our paper also acknowledges that RECIST scoring system is not used in real world practices especially in liver transplant centers where LiRADS criteria is standard of care in assessing HCC treatment response.
We think this paper, which represents real-world practice parameters adds a new perspective and additional insight to HCC treatment response that is applicable to day-to-day clinical practice in liver transplant center.

Changes in the text:

Comment 15: The results of this manuscript are important but they do not shed new light on the topic. It is already known that patients with lower MELD, PVT, and CP tend to do worse after Y90 and that these factors need to be kept in mind. There is discussion on REILD but this is in general something that doesn't occur as frequently anymore now that Y90 has been fine tuned over time.
Reply 15: We appreciate the reviewer's feedback and agree that REILD is not as frequently seen given more careful evaluation and selection for Y90 during our multidisciplinary liver tumor board. This is also stated in our manuscript in lines 352-354. However, our study does capture patients treated for HCC as far back as 2008 when REILD was likely more frequent. We have also added that REILD was not the only cause of decompensation in our patients but

that general liver decompensation (not meeting REILD criteria but includes under the categories of Table 8 the development of worsening liver failure, encephalopathy, ascites/ fluid retention, etc) also influenced MELD and CP. We have added additional comments in the text.

Per our prior response in above questions, we feel our study contributes further by providing insight into real-world practice vs. strict parameters of clinical trials. As mentioned above, standard practice for HCC diagnosis, post treatment response involves utilization of the LiRADS. We have added into the manuscript further description regarding the differences of LiRADS and mRECIST and how there can be an advantage of LiRADS over mcRECIST. We acknowledge our study does not follow mRECIST criteria, but this again emphasizes the significant differences that can be seen between clinical trial and real-world practice, and how this can add further insight and improve our understand of HCC treatment management. We believe our small study represents a stepping stone in bridging the efficacy- effectiveness gap, thereby, providing the connection between efficacy outcomes and meaningful change in the real-world clinical context. Real-world data also captures a more diverse and real-world representation of patient groups that are also often under-represented patient populations in clinical trials. We have added additional comments to the limitations and strengths of our study based on this.

Changes in the text: Page 11-12, line 266-284; Page 16, lines 415-416.

Reviewer B

This data set has the potential to provide useful information in the field of HCC treatment. However, the current study is lacking in a number of areas. Primarily description of the cohort and focused well defined outcomes. Because of these issues it is difficult for the reader to determine the value of and how to interpret the results of this study. Although real-world experience is a useful addition to the literature and the volume of data presented is admirable, the way in which it is presented and lack of context limits its utility.

Specific issues:

Comment 1: "One of the primary issues is lack of detail regarding the study population: Only a cursory description of how many received other LRT. It should be detailed how many were treated prior to and after y90 with other forms of LRT rather than just if it was given or not. Also, no information on how many received more than 1 y90 treatment and how they were treated in the data if they had two treatments."

Reply 1: We appreciate the Reviewer's feedback regarding details of Y 90 therapy, other forms of LRT and the timing and order of each treatment. Due to the heterogeneity of our data, we are not able to describe how many patients were treated prior or before Y90, or how many received more than one Y90. We acknowledge the lack of granularity of details is a weakness to our study and if we put this in our limitations section.

However, we have provided more granular data to Table 1 that was easily accessible in our database. This includes cancer staging, tumor location (unilobar/ bilobar/ diffuse), PVT (bland vs. malignant)/ (right/ left/ main branch), treatment with Y90 with single or bilobar, # of patients who received transplant, # of patients who were within Milan criteria.

Changes in the text: Page 17, lines 425-427

Comment 2: "Inadequate description of tumor stage. There is no mention of how many patients were within Milan criteria. There was no mention of BCLC stage. There was also no description of multifocal vs unifocal disease or size of largest tumor."
Reply 2: We have added details of Milan criteria, multifocal vs. unifocal disease to Table 1.

Unfortunately, we were not able to capture tumor size in our data. This has been acknowledged as a limitation in our paper. However, we have placed cancer staging details to Table 1 as well. We acknowledge that we did not use BCLC staging at the time of our study. This again reflects real world practice in the US where BCLC is generally not used. This also has been placed in our limitations section.

Changes in the text: Page 7, lines 165-166; Page 18, lines 436-439

Comment 3: "There is no mention of how many patients underwent liver transplant, if any. There is mention that they were censored, but no distinction between censoring because of transplant or for other reasons. It is difficult to contextualize the results without a clear picture of how many were transplant candidates initially, how many progressed outside of transplant criteria, and who many underwent successful transplant."

Reply 3: We have added the # of patients transplanted in Table 1.
We acknowledge that we do not have the granular details available regarding patients who were initially transplant candidates who then progressed outside of transplant criteria.

Comment 4: "Results: The introduction describes evaluating OS, PFS, and ORR as primary outcomes and complications as secondary outcomes. However, in the results the cohort is broken into a number of different groups to try to determine if certain factors (MELD, CP, TB) are associated with survival. I feel that if the purpose of the study is descriptive, most of the attempts to determine which factors correlate with outcomes should be eliminated. This would probably be a separate study and require larger sample size. In addition, most of these

factors (MELD etc) are already associated with worse outcomes, even without y90, so this seems redundant."

Reply 4: We appreciate the Reviewer's feedback regarding this. We put these detailed variables in the study as a way to provide information in a real-word setting. We definitely agree that there is clear contribution of OS, PFS, etc from underlying liver function.

Changes in the text:

Comment 5: The tables could be consolidated and improved for clarity (e.g. table 2 – what does "failed" signify, death?).

Reply 5: "Failed" signified death or progression. OS censored means the patient is still alive. PFS censored means cancer has not progressed. We have put a footnote under Table 2 for clarity.

Changes in the text:

Comment 6: Table 5 appears to show worsening of MELD over time. Hard to know how to interpret the influence of y90 since patients with cirrhosis and HCC tend to deteriorate over time without curative treatment (transplant).

Reply 6: We appreciate the Reviewer's comment regarding this. To provide further clarification, transplanted patients were censored from this data. With the improvement in MELD scores, what we found were that patients with intact liver function had longer survival. Patients with inferior liver function did not survive long enough, thus, the MELD scores improved.
(Given the additional tables made for revision, Table 5 is now Table 6 in our updated manuscript)

Changes in the text:

Comment 7: Discussion: The authors state that there was an increased median OS in CP A compared to B, but subsequently say it is not statistically significant in the results. This is phrased in a misleading way. In addition, later in the discussion the authors state that "In our multivariate analysis, older age, higher cancer stage, higher MELD and CP class, and presence of PVT correlated with significantly lower OS." These seem to be contradictory.

Reply 7: We appreciate the Reviewer's comments. To provide further clarity under the Discussion next to that statement, we have added that although Childs A did have longer median survival than Childs B, the difference in survival between the groups were not significantly different.

Regarding our multivariate analysis that correlated older age, higher cancer stage, higher MELD and CP, etc with lower OS, we have made revisions and placed this statement earlier on discussion to avoid the perception of contradiction.

Changes in the text: Page 10-11, line 239-248.

Comment 8: Regarding the difference in survival with PVT vs no PVT. Consider evaluating extent of clot burden as a potential difference between your experience and that of Mazzaferro cited in the discussion. Clot burden has been associated with prognosis. Could also consider describing if patients were anti-coagulated or not.
Reply 8: We appreciate the Reviewer's comments regarding this question. We agree that when PV is involved, depending on the extent of clot, this can play a role in survival.
In general, for tumor thrombus causing PVT, it is not our standard practice to initiate anticoagulation. In standard practice, we would only do so for bland thrombus affecting the main PV for patients awaiting liver transplant.

We have also added to Table 1 data regarding PVT(extent, bland/ malignant)

Changes in the text:

Comment 9: There is mention of mRECIST in the discussion, but not data on this is given.
Reply 9: The Reviewer's bring up an important point of discussion regarding our lack of use in mRECIST in our study parameters. To further explain our reasoning, our institution is a very active liver transplant center, and our radiologists utilize the Liver Imaging Reporting Data Systems (LiRADS) criteria for diagnosis of HCC as well as response to treatment. This became standard of care and practice in the United States in 2011 when the American College of Radiology launched LiRADS. This was implemented to standardize interpretation and reporting of diagnosis of HCC lesions as well as assessing the response to HCC lesions after treatment with locoregional therapy.

We acknowledge that mRECIST is used as surrogate endpoints in clinical trials. However, all liver transplant centers in the United States use LI-RADS, and LI-RADS treatment response (LR-TR) algorithm extends the definition of viable tumor by adding new imaging features such as washout appearance and enhancement similar to pretreatment, whereas the mRECIST considers arterial phase hyperenhancement (APHE) as the only characteristic of a viable tumor.

Studies have compared the diagnostic performance between the LiRADS algorithm and mRECIST criteria with reported results varying considerably. One recent meta-analysis comparing mRECIST and LiRADS criteria have suggested that LiRADS criteria have better specificity than mRECIST, without a significant difference in sensitivity for the diagnosis of pathologically viable HCC after LRT. (Kim DH et al. LI-RADS Treatment Response versus Modified RECIST for Diagnosing Viable Hepatocellular Carcinoma after Locoregional Therapy: A Systematic Review and Meta-Analysis of Comparative Studies. Taehan Yongsang Uihakhoe Chi. 2022 Mar;83(2):331-343.)

Our paper acknowledges that RECIST scoring system is not used in real world practices especially in the US where liver transplant centers use LiRADS criteria as standard of care in assessing HCC treatment response. Our study represents real-world practice parameters and adds a new perspective and additional insight to HCC treatment response that are applicable to day to day clinical practice in liver transplant centers. Because our study is focused on real-world practices using LiRADS , our study focused on actual clinical decision making rather than traditional stable disease criteria used in clinical trials,which can inflate durability of benefit. This makes our study a potential strength.

We have added this explanation to the manuscript.
Changes in the text: Pages 11-12, lines 266-290.