PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	The association between Tendon Ruptures and Use of Fluoroquinolone, and other oral antibiotics: A 10-year Retrospective Study of 1 million U.S. senior Medicare beneficiaries
AUTHORS	Baik, Seo; Lau, Jason; Huser, Vojtech; McDonald, Clement

VERSION 1 – REVIEW

REVIEWER	Mahyar Etminan University of British Columbia, Vancouver, Canada
REVIEW RETURNED	15-Oct-2019

GENERAL COMMENTS	This study attempts to show that previous studies that have shown a link to an increase in risk of tendon rupture (TR) and related events are possibly biased due to confounding by indication. My main concerns is with the design mainly with the cohort description and FQ ascertainment in the study. The authors state "Patients became eligible for "the study" at their Medicare enrolment but prescription data did not become available until their Part D enrolment. We followed them from their entry in Part D (while accounting for left truncation[24]) until their death, switch to a capitated plan, disenrollment from Medicare or 12/31/2016 – whichever came first". When reading this statement I can't exclude the possibility that the person time for a patient who survived to receive a FQ prescription in part D is misclassified as exposed when in reality this is unexposed person time leading to immortal time bias. Also, a patient included in the study who has received a FQ and has not experienced an event would be a prevalent user. Prevalent user bias leads to null associations. Authors must have a more clear and robust description of a new user design for this study. Moreover, it is not known when with respect to the FQ users when the index date (cohort entry date) for the comparator antibiotics was chosen. The same date as the FQ index date?. There should be a clearly defined look back period that can identify new users of all study drugs with a clearly defined index date. -The type of variables included in the model should be carefully revised. Some of these variables are not really confounders for this specific topic. For example, why are variables MI, chronic kidney disease, atrial fibrillation protective for TR? (Table 2) I cant think of any biologically plausible explanation. Table 2 should have a row that compares all FQs as a class vs Amox and Azithromycin (separately). Also the categories that compare the risk with different types of FQs (Cipro, Levo, Moxi)

	should be devoid of all FQs from the comparator/reference group. For example, the comparator group for moxifloxacin should not include any levo or cipro users. This might dilute the HRs. In Table 4 the HRs for FQ vs Amoxi and FQ vs Azithromycin for the Achilles tendon analysis have a very wide confidence interval suggestive of low events with the upper bound that doesn't exclude a harmful association. This might suggest lack of power for this analysis and all others analyses with wide confidence intervals.

REVIEWER	Charles Bennett WJB Dorn VA
REVIEW RETURNED	11-Dec-2019

GENERAL COMMENTS	The paper address TR and FQ use. While the stats indeed do not support a causal relationship or an association, this is the case with many known serious ADRs. No stats were found for numerous other clinically relevant sADRs including epoetin-associated pure red cell aplasia (NEJM), clopidogrel-associated TTP (NEJM), epoetin-associated VTE (JAMA), and rituximab-associated PML (Blood). Stats may not be the best way to find these sADRs. Did the authors look at the FAERS reports- as done by Raisch et al. What did this show?

VERSION 1 – AUTHOR RESPONSE

Reviewer(s)' Comments to Author:

Reviewer: 1

Reviewer Name: Mahyar Etminan

Institution and Country: University of British Columbia, Vancouver, Canada

Please state any competing interests or state ‘None declared’: None

Please leave your comments for the authors below This study attempts to show that previous studies that have shown a link to an increase in risk of tendon rupture (TR) and related events are possibly biased due to confounding by indication.

1. My main concerns is with the design mainly with the cohort description and FQ ascertainment in the study. The authors state “Patients became eligible for “the study” at their Medicare enrolment but prescription data did not become available until their Part D enrolment.’ We followed them from their entry in Part D (while accounting for left truncation [24]) until their death, switch to a capitated plan, or disenrollment from Medicare or 12/31/2016 – whichever came first”.

When reading this statement I can't exclude the possibility that the person time for a patient who survived to receive a FQ prescription in Part D is misclassified as exposed when in reality this is unexposed person time leading to immortal time bias.
The median number of days between Medicare entry and Part D entry is "zero" days. The inter quartile range is 0–122 days long. We agree with the reviewer that inclusion of such lag could have result in biased estimate of TR risk. However, we addressed this potential problem in the submitted paper by leaving all patient's time before Part D enrollment out of the at-risk set, until they were enrolled in Part D, an approach called survival analysis with delayed entry or left truncation (as asserted in the Method section). By doing so, we avoided bias in TR risk estimates of concern to the reviewer. Furthermore, all covariates but gender, race and rural residency in our survival analysis are time-varying which provides general protection against immortal time bias.

2. Also, a patient included in the study who has received a FQ and has not experienced an event would be a prevalent user. Prevalent user bias leads to null associations. Authors must have a more clear and robust description of a new user design for this study. Moreover, it is not known when with respect to the FQ users when the index date (cohort entry date) for the comparator antibiotics was chosen. There should be a clearly defined look back period that can identify new users of all study drugs with a clearly defined index date.
The reviewers point is relevant. We had not obtained a new user design (incident cohort) by washing out prevalent antibiotic users in the submitted paper, Instead, we washed out the event of interest (TRs) occurred in the first year of the study entry. We did this to increase the likelihood that we were looking at the first tendon rupture to be sure we had time relation between TRs and antibiotic use right . For the TR washout, we excluded all individuals with less than 1-year of follow-up after the study entry to ensure comparable data for the TR and no TR groups.

Initially, we did not think it was necessary to employ a new user design regarding the study antibiotics because they are used for such short periods for acute infections and because studies suggesting a relationship between FQ use and TR assumed the effect was fairly immediate occurring from 30-60 days of antibiotic use; so we focused on this time window. Table 1 below in this letter presents the distribution of days of supply for each antibiotic prescription. The most frequent days of supply (i.e., mode) for each antibiotics was pretty short 5-10 days.

Table 1. Distribution of Days of Supply by Antibiotics

		Days of Supply					
Antibiotics	**N of Prescription**	**Mean**	**Standard Deviation**	**Mode**	**1st quartile**	**Median**	**3rd quartile**
AMX	2,113,929	7.76	5.95	10	5	7	10
AMC	1,144,887	9.77	4.11	10	7	10	10
AZM	2,550,160	5.96	6.44	5	5	5	5
LEX	1,304,367	9.55	9.63	10	7	7	10

CIP	2,023,911	9.03	7.07	7	5	7	10
LVX	1,073,824	8.63	4.64	10	7	7	10
MXF	320,454	13.33	7.49	10	8	11	15

Incident user designs are especially important for evaluating the long-term drug effect on chronic conditions. We were not as sure they would be important for studies of short-term antibiotic use because individuals in the incident cohort would not be really “new” users.

However, they are usually considered to be a stronger design, so we did implement a new user design. We created a 3-month washout cohort that excluded any patients who was prescribed any study antibiotics during their first 3 months after Part D enrollment, and ignored the data for patients not taking study antibiotics over that same time period. By doing this, we could minimize prevalence bias and assure that subject were starting a new prescription at least 3 months after any previous antibiotic and also beyond the 2-month post antibiotic window which most of the risk of TRs has been presumed to pass. We present the results of this new user design in the current version of our paper, but the results were almost identical to the results with prevalent users (Table 2 below).

Table 2. Comparison of HRs for any Tendon Rupture between Prevalent User Design and Incident User Design.

		Prevalent Cohort	**3-Month Washout Cohort**
	Temporal Exposure	**HR(95% CI)**	**HR(95% CI)**
AMX VS. NO AMX	**≤ 30 days**	0.87(0.82,0.92)‡	0.86(0.80,0.92)‡
	31 – 60 days	0.92(0.86,0.98)†	0.94(0.87,1.01)
	≥ 61 days	1.00(0.99,1.02)	1.00(0.98,1.02)
AMC VS. NO AMC	**≤ 30 days**	0.97(0.90,1.04)	0.93(0.85,1.02)
	31 – 60 days	1.04(0.96,1.14)	0.95(0.85,1.05)
	≥ 61 days	1.06(1.04,1.08)‡	1.07(1.04,1.09)‡
AZM VS. NO AZM	**≤ 30 days**	1.02(0.97,1.08)	0.99(0.93,1.06)
	31 – 60 days	0.94(0.88,1.00)†	0.90(0.84,0.98)†
	≥ 61 days	1.07(1.06,1.09)‡	1.07(1.05,1.09)‡
LEX VS. NO LEX	**≤ 30 days**	1.26(1.18,1.34)‡	1.31(1.22,1.41)‡
	31 – 60 days	1.09(1.01,1.18)†	1.05(0.95,1.15)

	≥ 61 days	1.09(1.07,1.12)‡	1.08(1.05,1.11)‡
LVX VS. NO LVX	**≤ 30 days**	1.15(1.07,1.23)‡	1.14(1.05,1.25)†
	31 – 60 days	1.06(0.97,1.16)	1.09(0.98,1.21)
	≥ 61 days	1.00(0.97,1.02)	1.02(1.00,1.05)
CIP VS. NO CIP	**≤ 30 days**	0.94(0.88,1.00)†	0.96(0.89,1.03)
	31 – 60 days	0.95(0.89,1.02)	0.92(0.85,1.01)
	≥ 61 days	0.96(0.94,0.98)‡	0.96(0.94,0.98)‡
MXF VS. NO MXF	**≤ 30 days**	0.69(0.49,0.97)†	0.59(0.37,0.93)
	31 – 60 days	0.69(0.46,1.02)	0.71(0.43,1.15)
	≥ 61 days	1.00(0.95,1.05)	0.99(0.93,1.06)

In a survival regression analysis, the index date is the time the event of interest (e.g., death, or the first occurrence of a TR). The survival regression analysis takes each patient who experienced an event, makes that the "index time" and compares that patients characteristics at that time with the characteristics of all other patients in the risk set at the same time. It then takes that patient out of the risk set and then repeats the process successively for every patient who experienced an event. This is a very powerful method.

2. The type of variables included in the model should be carefully revised. Some of these variables are not really confounders for this specific topic. For example, why are variables MI, chronic kidney disease, and atrial fibrillation protective for TR? (Table 2) I can't think of any biologically plausible explanation.

   We included all of the Medicare chronic diseases as general adjuster for disease burden. We did not want to pick and choose among them based on our subjective biases. Furthermore, 1) we were using a competing risk regression analysis in which death was the competing risk and many of the chronic diseases, especially the 3 singled out by the reviewer were quite relevant to the prediction of death. 2) These chronic conditions are also likely differentially associated with the use of study antibiotics and also relevant for that reason.

   Finally, as the reviewer pointed out, the three selected conditions were all predictive but of reduced, not an increased, risk of Tendon Ruptures (TRs). That makes sense because the named conditions would be likely to reduce amount of physical activity that would might otherwise lead to TRs. These also are associated with an increased the death rate and thus shrinking the exposure time during which a TR could occur. We see the same protective

pattern with all of the life threatening conditions in our analysis. The “protective” effect of lung cancer for any TR has especially large, as was over the top 6 fold increase in death risk.

To be sure that we did not distort the results by including these three conditions, we re-ran the analyses without them. The results were unchanged (See Table 3 in this letter below).

Table 3. Comparison of HRs with and without 3 conditions (MI, chronic kidney disease, and atrial fibrillation)

		With 3 Conditions	Without 3 Conditions
Comparison	**Temporal Exposure**	**HR(95% CI)**	**HR(95% CI)**
CIP VS. LVX	**≤ 30 days**	0.82(0.74,0.90)	0.82(0.74,0.90)
CIP VS. MXF	**≤ 30 days**	1.36(0.96,1.92)	1.36(0.96,1.91)
LVX VS. MXF	**≤ 30 days**	1.66(1.18,2.35)	1.66(1.18,2.35)
CIP VS. A3MX	**≤ 30 days**	1.08(0.99,1.17)	1.07(0.99,1.17)
CIP VS. AZM	**≤ 30 days**	0.92(0.85,1.00)	0.91(0.84,0.99)
CIP VS. LEX	**≤ 30 days**	0.74(0.68,0.81)	0.74(0.68,0.81)
CIP VS. AMC	**≤ 30 days**	0.97(0.88,1.07)	0.97(0.88,1.07)
LVX VS. AMX	**≤ 30 days**	1.32(1.20,1.45)	1.32(1.20,1.44)
LVX VS. AZM	**≤ 30 days**	1.13(1.03,1.23)	1.12(1.02,1.23)
LVX VS. LEX	**≤ 30 days**	0.91(0.83,1.00)	0.91(0.83,1.00)
LVX VS. AMC	**≤ 30 days**	1.19(1.07,1.32)	1.19(1.07,1.32)
MXF VS. AMX	**≤ 30 days**	0.79(0.56,1.12)	0.79(0.56,1.12)
MXF VS. AZM	**≤ 30 days**	0.68(0.48,0.95)	0.67(0.48,0.95)
MXF VS. LEX	**≤ 30 days**	0.55(0.39,0.77)	0.55(0.39,0.77)
MXF VS. AMC	**≤ 30 days**	0.71(0.51,1.01)	0.72(0.51,1.01)
FLQ VS. AMX	**≤ 30 days**	1.04(0.91,1.18)	1.04(0.91,1.18)
FLQ VS. AZM	**≤ 30 days**	0.89(0.78,1.01)	0.88(0.78,1.01)
FLQ VS. LEX	**≤ 30 days**	0.72(0.63,0.82)	0.72(0.63,0.82)
FLQ VS. AMC	**≤ 30 days**	0.94(0.82,1.08)	0.94(0.82,1.08)

3. Table 2 in the paper should have a row that compares all FQs as a class vs Amoxicillin and Azithromycin (separately).

   We assume the reviewer meant Table 3 not 2, because all such comparisons in Table 3 would fit directly. So, we included comparisons of all FQs as a class against all of the non FQ antibiotics (including Amoxicillin and Azithromycin separately) in Table 3.

4. Also the categories that compare the risk with different types of FQs (ciprofloxacin, levofloxacin, and moxifloxacin) should be devoid of all FQs from the comparator/reference group. For example, the comparator group for moxifloxacin should not include any levofloxacin or ciprofloxacin users. This might dilute the HRs.

This suggestion collides with the dictum from our survival analysis consultant from Mayo who says we should never manipulate time series data in any way that looks into the future. It makes the analysis very prone to immortal time bias. The creation of the comparator group would require searching forward ahead through years of prescription data to find those that had one kind of antibiotic and not another type and removing patients in the second category and would violate that dictum.

More importantly, we treat all of the covariates except gender, race and rural residency as time varying covariates and drug groups in our study population are not mutually exclusive. Patients can be on multiple antibiotic classes at any one point in time epoch. About 2.5% patients took more than one antibiotics simultaneously. Our survival regression analysis delivers the independent effect of each of the study antibiotics by cancelling out effects of other antibiotics in the calculation of hazard ratio. For example, in the estimation of relative TR risk for ciprofloxacin, both ciprofloxacin group and comparator group (reference: no ciprofloxacin users) could be on other FQs and/or non-FQ antibiotics. However, in the calculation of HR for ciprofloxacin, effects of other FQs and non-FQ antibiotics are cancelled out as seen in equation below.

HR=(h_0 (t)exp(β_1 CIP+β_2 LVX+β_3 MXF+β_4 AMX+β_5 AZM+β_6 LEX+Bcovariates))/(h_0 (t)exp(β_2 LVX+β_3 MXF+β_4 AMX+β_5 AZM+β_6 LEX+Bcovariates))=exp(β_1)

So our reported HR for each antibiotics cannot be distorted or diluted.

In Table 4 the HRs for FQ vs Amoxicillin and FQ vs Azithromycin for the Achilles tendon analysis have a very wide confidence interval suggestive of low events with the upper bound that doesn't exclude a harmful association. This might suggest lack of power for this analysis and all others analyses with wide confidence interval.
We agree the confidence intervals are indeed wide for AZM and AMX (as well as for AMC). Further, Achilles tendon ruptures were rare in our study population—around 2.5% of the population, influencing statistical power. The statistical power was further reduced by dividing drug exposure into three sub-groups based on the relationship between antibiotic exposure and

TR (i.e., ≤30 days, 31-60 days, ≥61 days). In the original report, cells in Table 4 only reported the data for one of these windows, HR for TRs within 30 days of antibiotic exposure.

The last line in the results section of our original submission said “When comparing the risk of FQs as a class against that of non-FQ antibiotics, most of the non-FQ antibiotics had significantly greater risk than the FQ class as a whole across all TR sites.” We believe this is statement triggered the appropriate concern by the reviewer, because the submitted version carried no data to support this statement. We wrote that line while reading data in our summary spreadsheet that showed the comparison of FQs with each of all other antibiotics from the 3 time windows combined, but we mistakenly failed to include that data in Table 4 of our original submission.

In the revised version, we added these 4 missing rows to the bottom of Table 4. As you can see with the data from all the time windows combined, FQs as a group have significantly less risk of TR than each of non-FQ antibiotics for Achilles tendon rupture, and is true for other types of FQs.

Reviewer: 2
Reviewer Name: Charles Bennett
Institution and Country: WJB Dorn VA
Please state any competing interests or state 'None declared': none

Please leave your comments for the authors below.
1. The paper address TR and FQ use. While the stats indeed do not support a causal relationship or an association, this is the case with many known serious ADRs. No stats were found for numerous other clinically relevant sADRs including Epoetin associated pure red cell aplasia (NEJM), clopidogrel- associated TTP (NEJM), epoetin-associated VTE (JAMA) and rituximab-associated PML (Blood). Stats may not be the best way to find these sADRs. Did the authors look at the FAERS reports- as done by Raisch et al? What did this show?
We are not certain what Dr. Bennet means by "stats" but presume he means statistical methods. If so, we did not mean to argue that statistical methods are supreme or the only valid methods for discovering ADRs. Further, we admire the clever investigations that found the many ADRs cited in his 4 references. Some of the ADRs were very rare. We believe there were only 11 available cases for clopidogrel-associated TTP, and they assembled only 175 cases of pure red cell aplasia associated with epoetin and 52 with leukoencephalopathy. The relation between these 3 drugs and the ADR all have well known causal mechanisms, immunologic for the first two and prion infections for the last one. These ADRs also have specific tests or outcomes that prove the "cause". The situation regarding TRs is quite different. There is no clear biologic mechanism to explain how FQs would cause tendon ruptures and no confirmatory test that can verify that a given tendon rupture is due to FQ use.

Furthermore, contrary to what might be implied by "While the stats indeed do not support a causal relationship or an association", our stats do support significant associations between some of the antibiotics and TRs; positive associations (increased risk) for TR with the use of cephalexin or levofloxacin and negative association with use of ciprofloxacin and moxifloxacin. Furthermore, ciprofloxacin represents more than half of the FQ users and thus provides statistical power to detect risk reduction.

Our data collide with the assertion that the risk of FQs on TRs is a class effect. On the one hand we saw a strong association between cephalexin use and the risk of TRs equal to or greater than levofloxacin. Cephalexin is not a FQ, and it lacks the stimulatory effect on metalloprotease that some assume is the causal mechanism for FQ associated TRs. On the other hand, we saw no relationship between ciprofloxacin and TRs though ciprofloxacin was the most frequently used FQ in our study and it has a strong activating effect on metalloprotease activity. Together these two facts undercut the theory that the FQs cause TRs via their effect on metalloproteases.

We did look at the FAERS reports- as done by Raisch et al who is a coauthor of one of reviewer's cited references. We respect the work he has done, the utility for the FEARS database and the EBGM method for scoring associations found in FAERS.

The FAERS report the reviewer mentioned includes a very high EBGM score for the ratio of TR events associated with FQs compared to other drugs. However, in general, adverse events are under reported in spontaneous reporting system like FAERS and the FAERS database is still not appropriate for estimating incidence rate because of lack of a true denominator for the population at risk. The FAERs data base has been very useful, but results based upon FAERS database are not guaranteed to be correct. They can be biased by several factors, association of a drug with an adverse effect might be explained by patient's conditions (indication bias) and other drugs which are

often co-administered (confounding bias). DuMouchel (and his co investigators) who is associated with the development of the EBGM and very active in this fields described logistic methods as being [too] computational intensive. But was open to them.

“Many patients, particularly the elderly, take more than two medications. Searching for and analyzing, the effects of polypharmacy to find higher order interactions is a challenge that can be approached with regression-based strategies”.
Out study worked with elderly patients who take many medications and we did use a regression method which has important advantages of the simpler ADR methods. We do not claim that one method has a monopoly, and truth is more likely to arise from multiple independent approaches. We believe that multiple approaches are good for discovering truth.

Putter H, Fiocco M, Geskus RB. Tutorial in biostatistics: competing risks and multi-state models. Stat Med 2007;26:2389–430. doi:10.1002/sim.2712
Suissa S. Immortal time bias in observational studies of drug effects. Pharmacoepidemiology and drug safety. 2007 Mar;16(3):241-9.

Suissa S. Immortal time bias in observational studies of drug effects. Pharmacoepidemiology and drug safety. 2007 Mar;16(3):241-9.

Bennett CL, Luminari S, Nissenson AR, Tallman MS999, Klinge SA, McWilliams N, McKoy JM, Kim B, Lyons EA, Trifilio SM, Raisch DW. Pure red-cell aplasia and epoet in therapy. New England Journal of Medicine. 2004 Sep 30;351(14):1403-8.

Bennett CL, Connors JM, Carwile JM, Moake JL, Bell WR, Tarantolo SR, McCarthy LJ, Sarode R, Hatfield AJ, Feldman MD, Davidson CJ. Thrombotic thrombocytopenic purpura associated with clopidogrel. New England Journal of Medicine. 2000 Jun 15;342(24):1773-7.
Bennett CL, Silver SM, Djulbegovic B, et al. Venous Thromboembolism and Mortality Associated With Recombinant Erythropoietin and Darbepoetin Administration for the Treatment of Cancer-Associated Anemia. JAMA. 2008;299(8):914–924. doi:https://doi.org/10.1001/jama.299.8.914
Carson KR, Evens AM, Richey EA, Habermann TM, Focosi D, Seymour JF, Laubach J, Bawn SD, Gordon LI, Winter JN, Furman RR. Progressive multifocal leukoencephalopathy after rituximab therapy in HIV-negative patients: a report of 57 cases from the Research on Adverse Drug Events and Reports project. Blood. 2009 May 14;113(20):4834-40.

Arabyat RM, Raisch DW, McKoy JM, Bennett CL. Fluoroquinolone-associated tendon-rupture: a summary of reports in the Food and Drug Administration’s adverse event reporting system. Expert opinion on drug safety. 2015 Nov 2;14(11):1653-60.
Sakaeda T, Tamon A, Kadoyama K, et al. Data mining of the public version of the FDA Adverse Event Reporting System. Int J Med Sci 2013;10:796–803.
Almenoff JS, Pattishall EN, Gibbs TG, DuMouchel W, Evans SJ, Yuen N. Novel statistical tools for monitoring the safety of marketed drugs. Clinical Pharmacology & Therapeutics. 2007 Aug;82(2):157-66.

VERSION 2 – REVIEW

REVIEWER	Charles. Bennett U of. South. Carolina United. States
REVIEW RETURNED	09-Feb-2020

GENERAL COMMENTS	The. Paper. Does not reference. The. 2006 Paper does not reference the 2006 and 2008 citizen petitions on the subject The paper does not. Include. The. Long. History of tendon ruptures. And. Fq. Nor the long litigation on this. Topic. FDA. Reviewed the data nicely in 2015. Please add this. Material. To. The. Text. The fda conclusion is in disagreement. With yours. This must be discussed.

REVIEWER	Joseph Nolan Northern Kentucky University
REVIEW RETURNED	25-Jun-2020

GENERAL COMMENTS	This study is complex and one general comment I have is that I don't believe it repeatable based on the description provided in methods. Based on the results shown (but lacking clarity in how some of the modeling was done) the primary conclusions within the abstract seem reasonably believable. I indicate "minor revision" specifically because I am not convinced that this necessarily requires re-analysis (see notes on interaction below) but rather would guess that the methodology might just need better explanation. Depending on answers to questions I pose below, however, it could be "major". I would assert that before publication, the following statistical concerns should be addressed: 1. The study is described as prospective, a word that generally suggests that one is watching over time for outcomes to occur, and suggests some level of experimental design. I don't believe any of that is present here. Lacking any access to the participants and having only access to deidentified Medicare data, I cannot see how it can truly be anything other than retrospective. If subjects are actually followed in some way, the methodology needs to be better described. Otherwise, the study should not be described as prospective. 2. The analysis methodologies (logistic regression, survival analysis) applied seem reasonable. However the number of predictors being considered is quite large, and many of them are likely to be correlated and might also have relevant interactions. The model(s?) being considered attempt to account for all of the variables (40+) in Table 2 as well as the antibiotics in Table 3. The amount of intercorrelation among predictors is unclear, but seems likely to confound results. Additionally it is unclear whether or not statistical interaction has been considered. 3. It is unclear what the results in Table 2 represent. One model with 40+ variables? 40+ models with one variable each? Are interactions among predictors considered? The model(s?) represented by Table 3 is equally unclear. How these model(s?) are constructed needs substantially better explanation. In particular, it seems like there would be potential interactions among antibiotics that may not have been modeled.

	4. I would point out that the only reasonably interpretable information in Tables 2 & 3 are the 95% CI's, so that aspect is quite important. The large sample size is relevant to these being so potentially useful. However, there are 132 CI's found in Table 3, meaning that if no adjustment for multiplicity is made, we would fully expect that 7 or 8 of them would come up as statistically significant (CI not containing 1.000) by random chance. Further, hazard ratios alone do not indicate actual relevance (or lack thereof). Based on the 34K TR that were reported in only about 3.4% of study cases. Some idea of relevance needs to be given in terms of absolute risk. If the risk goes from 3.4% maybe 3.6%, probably no one cares. That isn't addressed by looking only at hazard ratios. On this same note, there are many percentage numbers being reported in the text of the paper and I am unable to verify anything in terms of how those were derived (for example "LVX is the only FQ to exhibit a significant increase in TR risk – of 16%, and 120% for rupture of rotator cuff and Achilles TR respectively in the <30 day window" – what did these numbers come from and what do they mean?).

VERSION 2 – AUTHOR RESPONSE

Reviewer(s)' Comments to Author:

Reviewer: 2

Reviewer Name: Charles. Bennett

Institution and Country:

U of. South. Carolina

United. States

Please state any competing interests or state 'None declared': None

Please leave your comments for the authors below The. Paper. Does not reference. The. 2006 Paper does not reference the 2006 and 2008 citizen petitions on the subject

The paper does not. Include. The. Long. History of tendon ruptures. And. Fq. Nor the long litigation on this. Topic. FDA. Reviewed the data nicely in 2015. Please add this. Material. To. The. Text. The fda conclusion is in disagreement. With yours. This must be discussed.

 This comment was a bit telegraphic and hard to follow. We already have included three references to the FDA's black box response to the use of FQs in the introduction of the current paper. The cited FDA references link to other sources. So it is not clear how many documents from the FDA that we should cite. We had to dig to find it, but I believe the 2015 report the reviewer referred to in the comment is a paper on which he was the senior author. We added it as a reference in a further discussion about the FDA considerations (pages 15-16).

Please note we wrote very long response to Dr. Bennett comments in the first round of this review including a fairly thorough discussion of the FDA's analysis of spontaneous reports of adverse and their EBGM method. We assume that that these earlier responses will be considered by the editor.

Reviewer: 3

Reviewer Name: Joseph Nolan

Institution and Country: Northern Kentucky University Please state any competing interests or state 'None declared': None Declared

Please leave your comments for the authors below This study is complex and one general comment I have is that I don't believe it repeatable based on the description provided in methods. Based on the results shown (but lacking clarity in how some of the modeling was done) the primary conclusions within the abstract seem reasonably believable. I indicate "minor revision" specifically because I am not convinced that this necessarily requires re-analysis (see notes on interaction below) but rather would guess that the methodology might just need better explanation. Depending on answers to questions I pose below, however, it could be "major". I would assert that before publication, the following statistical concerns should be addressed:

1. The study is described as prospective, a word that generally suggests that one is watching over time for outcomes to occur, and suggests some level of experimental design. I don't believe any of that is present here. Lacking any access to the participants and having only access to de-identified Medicare data, I cannot see how it can truly be anything other than retrospective. If subjects are actually followed in some way, the methodology needs to be better described. Otherwise, the study should not be described as prospective.

 The reviewer is right. So we replaced "prospective" with "retrospective".

2. The analysis methodologies (logistic regression, survival analysis) applied seem reasonable. However the number of predictors being considered is quite large, and many of them are likely to be correlated and might also have relevant interactions. The model(s?r) being considered attempt to account for all of the variables (40+) in Table 2 as well as the antibiotics in Table 3. The amount of intercorrelation among predictors is unclear, but seems likely to confound results. Additionally it is unclear whether or not statistical interaction has been considered.

 Our analytic model does includes a quite large number (80) of covariates made up of 59 predictors and the use of 7 different study antibiotics included in 3 mutually exclusive usage timings in relation to the TR yielding 21 different covariates to reach a total of 80.

Some of our predictors such as MI (myocardial infarction) may look irrelevant for the risk of TR. Because the death prevalence exceeded the TR prevalence, we employed a competing risk regression with death as the competing risk. And many of the predictors such as myocardial infarction (MI) were relevant to the predication of death. Some of the predictors were also likely to be differentially associated with the use of study antibiotics and relevant for that reason.

The usual worry about too many variables is the risk of overfitting. However, we did not face that risk because of our huge sample size (> 1million) and very large number, 35,000, of TR events.

Regarding the question of interactions and intercorrelation among predictors, we were reluctant to include interactions (as many as 80x80=6400) in our regression, because the number additional covariates would challenge our large sample size in terms of over fitting. In order to evaluate the amount of intercorrelation, we calculated a correlation matrix of estimated regression coefficients from our competing risk regression analysis (SAS PROC PHREG) considering their time-varying nature. The correlation matrix can deliver information about the strength of all pairwise correlation and indicate the existence of a collinear relationship between two predictors. All correlations (except diagonal elements) were at or below 0.33, about 98.6% were below 0.1 and only 0.2% of the 6400 correlations, were between 0.2 and 0.33. The largest of the pairwise correlations was 0.33 indicating minimal bias due to intercorrelation. We have attached the matrix in case the reviewer would like to see them. And none of the correlation >0.2 applied to any study drug. We included the evaluation of intercorrelation among predictors and acknowledged the absence of interactions in our analyses as a limitation (Page 19).

If the reviewer was asking why we didn't use some stepwise elimination of variables. It was because some experts strongly opposes the use of either forward or reverse stepwise selection. They argue that "full model fits (that is, leaving all hypothesized variables in the model regardless of P-value) are frequently more discriminating than fits after screening for significance."

3. It is unclear what the results in Table 2 represent. One model with 40+ variables? 40+ models with one variable each? Are interactions among predictors considered? The model(s?) represented by Table 3 is equally unclear. How these model(s?) are constructed needs substantially better explanation. In particular, it seems like there would be potential interactions among antibiotics that may not have been modeled.

 We apologize for the lack of clarity in our first draft. The data for tables 2, 3 and 4 all comes from the same set of cox analyses. Table 2 shows the HRs for all of the non-drug predictors reported in our competing risk regressions. And Table 3 focuses on the HRs associated with the use of the study antibiotics from the same set of analyses. In Table 3, we show the antibiotic HRs broken down by type of TR in separate columns and the time relationship between the TR and the use of the respective antibiotics in separate rows e.g. use within 30 days of the TR, use within 31-60 days of the TR and use >60 days before the TR. In this revision, we have divided Table 3 into two sections for ease of reading. The first section presents the HRs for each of the 7 study antibiotic with no use of that antibiotic as the reference. The second section compares any FQ use with the use of each Non-FQ antibiotic as the reference. Table 4 shows pairwise comparison of the different antibiotic or antibiotic classes. Tables 3 and 4 now includes the HRs for death which we should have included in the first place because the analysis is a competing risk analyses. Again, we were reluctant to include even two-way interactions in the model due to possible overfitting. In the discussion (page 19), we acknowledged the possible interactions among different antibiotics and the absence of such consideration as a limitation in our analyses.

4. I would point out that the only reasonably interpretable information in Tables 2 & 3 are the 95% CI's, so that aspect is quite important. The large sample size is relevant to these being so potentially useful.

However, there are 132 CI's found in Table 3, meaning that if no adjustment for multiplicity is made, we would fully expect that 7 or 8 of them would come up as statistically significant (CI not containing 1.000) by random chance.

 In Table 3, we reported hazard ratios of each antibiotic and confidence intervals, which are exponentiated regression coefficients from four competing risk regression analyses (one for each of the 3 tendon rupture types and one for the any TR combined). So the last one is not an independent assessment. And the last 3 rows represent a different cut at the same data to highlight the comparison of FQs with non-FQs.

The familywise type 1 error rate (false positive) may increase as we include more covariates in the model. However, reducing (or adjusting) type 1 error for null association increase the type 2 error rate (false negative), consequently hurt statistical power of the analysis. Several studies have suggested that if the selected covariates do not exceed the sample size and they are all plausible pathways to the outcome of interest, multiple tests can be performed without adjusting type 1 error rate, i.e., the risk of false positive is handled fairly well

However, in order to comply the review's concern, we tested the significance of each antibiotic in Table 3 with multiplicity corrected p-value and results stayed the same. Now in Tables 3 and 4 (Pages 35-36), we present statistically significant HRs based upon multiplicity corrected p-values.

Further, hazard ratios alone do not indicate actual relevance (or lack thereof). Based on the 34K TR that were reported in only about 3.4% of study cases. Some idea of relevance needs to be given in terms of absolute risk. If the risk goes from 3.4% maybe 3.6%, probably no one cares. That isn't addressed by looking only at hazard ratios.

 Absolute risk refers to the simple event rate in a group of people who received an intervention. In Table 1, we separately reported proportion of TR and TR risk per 1000 person-years by each study antibiotics. Two measures (proportion and TR risk per 1000 person-years) were pretty similar across different antibiotics except Moxifloxacin. We used TR risks per 1000 person-years along with HRs to explain no increased risk of TRs for fluoroquinolone as a class.

We implemented a competing risk regression analysis because death (competing risk) precludes the occurrence of tendon rupture (primary event of interest). The competing risk regression is a semi parametric model. It leaves the baseline hazard rate unspecified and thus do not assume knowledge of absolute risk. We pointed out this reality in the process of the first round of reviews and received the following from the editor.

The editor had asked: Please include details of absolute risk in the abstract and throughout the results section, in addition to the hazard ratio data provided.

After a communication with Amy Branch-Hollis the assistant editor, explaining that obtaining absolute risks is very difficult when the analysis is based on a competing risk, she responded in an email as of Jan 08, stating that we could ignore that request.

The content of her email was as follows;

Dear Dr. McDonald,

Please accept my apologies for the delay in my response, I hope this email finds you well.

Thank you for providing the explanations regarding absolute risk calculations. After reading your email and looking into your manuscript again, we have decided that we no longer require you to provide these numbers. When responding to these requests in your rebuttal, you can explain that we said these requests were no longer necessary.

Please accept our apologies for the confusion and for any inconvenience caused. If you have any further questions, please do not hesitate to contact us,

On this same note, there are many percentage numbers being reported in the text of the paper and I am unable to verify anything in terms of how those were derived (for example “LVX is the only FQ to exhibit a significant increase in TR risk – of 16%, and 120% for rupture of rotator cuff and Achilles TR respectively in the <30 day window” – what did these numbers come from and what do they mean?).

 HRs in Table 3 can be interpreted as the chance of an event (e.g., TR) occurring in the treatment arm divided by the chance of the event in the control arm. Instead of reporting HR itself, we reported percent of risk change for antibiotic use versus not. For example, HR of Achilles tendon rupture for levofloxacin (LVX) was 2.20, indicating 120% increased risk of Achilles tendon rupture for those who were on LVX compared to those not on LVX.

Arabyat RM, Raisch DW, McKoy JM, Bennett CL. Fluoroquinolone-associated tendon-rupture: a summary of reports in the Food and Drug Administration’s adverse event reporting system. Expert opinion on drug safety. 2015 Nov 2;14(11):1653-60.

Harrell Jr FE, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Statistics in medicine. 1996 Feb 29;15(4):361-87.

Cook RJ, Farewell VT. Multiplicity Considerations in the Design and Analysis of Clinical Trials. J R Stat Soc Ser A (Statistics Soc 1996;159:93. doi:10.2307/2983471

Perneger T V. What’s wrong with Bonferroni adjustments. BMJ 1998;316.

Rothman KJ. No adjustments are needed for multiple comparisons. Epidemiology 1990;1:43–6. doi:10.1097/00001648-199001000-00010

How results are presented (2): risks, ratios, NNT and NNH. Prescriber 2007;18:21–6. doi:10.1002/psb.62

VERSION 3 – REVIEW

REVIEWER	Joseph Nolan Northern Kentucky University USA
REVIEW RETURNED	12-Aug-2020

GENERAL COMMENTS	First I apologize for taking so long to turn this around – been quite busy lately. The revisions have satisfied much of my concerns, however I do have two that remain. First, I thank the authors for pointing out that the absolute risk is reported in Table 1, which I suspect I may have missed the first time. However, I do believe this should be further emphasized and actually applied to the discussion. For example, tell the reader in real terms what is the gain by getting “almost 14% fewer hospitalizations and half of death rate per 1000 patient years”? How many more deaths per capita are represented by a “119% increased risk of death in a <30-day window”? Build these in together so that the reader can fully understand the relevance (or lack thereof). This is an appropriate compromise, of sorts, to the editor’s request to include absolute risks throughout. Building in a few examples where you move to absolute risk will help the reader do the same when it comes to the tables. I think this should be required, as otherwise people will assuredly misinterpret statistical significance to automatically imply clinical relevance. Second, I note that you have provided one reference that suggests throwing all factors in the model with no interaction and ignoring model selection. While I note that the author of that manuscript seems to have the appropriate training, I doubt that too many statisticians would agree with that strategy and myself would view that as something to consider but not something that should be taken as a gold standard. The fact that you have apparently not even attempted to consider two-factor interactions (I immediately would grant that you can’t possibly consider all interactions). As you point out, you have substantial amounts of data. It seems that cross-validation strategies might actually be the best here, and that you certainly could examine two-factor interactions and include relevant ones if you chose. I’m really on the fence about whether this is an absolute must, and will leave it to the editor to make a decision as to whether further modeling would be required or this deemed sufficient.

VERSION 3 – AUTHOR RESPONSE

Reviewer(s)' Comments to Author:

Reviewer: 3

Reviewer Name: Joseph Nolan

Institution and Country:

Northern Kentucky University

USA

Competing interests: None Declared

Please leave your comments for the authors below First I apologize for taking so long to turn this around – been quite busy lately. The revisions have satisfied much of my concerns, however I do have two that remain.

First, I thank the authors for pointing out that the absolute risk is reported in Table 1, which I suspect I may have missed the first time. However, I do believe this should be further emphasized and actually applied to the discussion. For example, tell the reader in real terms what is the gain by getting "almost 14% fewer hospitalizations and half of death rate per 1000 patient years"? How many more deaths per capita are represented by a "119% increased risk of death in a <30-day window"? Build these in together so that the reader can fully understand the relevance (or lack thereof). This is an appropriate compromise, of sorts, to the editor's request to include absolute risks throughout. Building in a few examples where you move to absolute risk will help the reader do the same when it comes to the tables. I think this should be required, as otherwise people will assuredly misinterpret statistical significance to automatically imply clinical relevance.

→ The reviewer is asking us to specify the "gain" in absolute terms of "getting" fewer hospitalizations and half the death rate per 1000 years. That is, of course, would be nice addition if we could add 'adjusted' absolute risks. But in general, it is difficult to accomplish with Cox regression and near impossible with competing risk analyses. In this study, we implemented a competing risk regression analysis because death was a competing risk, which precluded the occurrence of tendon rupture, the outcome of interest. The competing risk

regression is a semi-parametric model. It leaves the baseline hazard rate unspecified, which is a key to derive the adjusted absolute hazard rate.

We explained this problem to Amy Branch-Hollis, assistant editor, on Jan 08, 2002—she responded: "*After reading your email and looking into your manuscript again, we have decided that we no longer require you to provide these numbers [referring to the absolute hazard rate]. When responding to these requests in your rebuttal, you can explain that we said these requests were no longer necessary*"

Instead, we did report two *crude (or unadjusted)* absolute measures - the % (prevalence) of all subject who experienced TRs before death and who experienced death without TRs in Table 1 below (rearranged the Table 1 in the manuscript). They are, however, not adjusted by the many covariate and cannot detect close timing relationships between the short-term drug use and the events of interest. In page 17, we have added 'unadjusted' absolute risks of tendon ruptures and death per 1000 person-year along with HRs for the readers to understand better.

Table 1. Unadjusted Risks of TRs and Death broken down by each type of antibiotics.

	TR		Death	
Rx	**%**	**1000 person-years**	**%**	**1000 person-years**
AMX	**3.7**	**7.6**	**3.7**	**7.6**
AUG	**3.7**	**7.4**	**5.6**	**11.2**
AZT	**4.0**	**8.1**	**4.7**	**9.6**
LEX	**4.1**	**8.0**	**5.8**	**11.4**
CIP	**3.7**	**7.4**	**6.3**	**12.5**
LVX	**3.8**	**7.5**	**9.4**	**18.5**
MXF	**5.2**	**8.8**	**14.6**	**24.4**

Second, I note that you have provided one reference that suggests throwing all factors in the model with no interaction and ignoring model selection. While I note that the author of that manuscript seems to have the appropriate training, I doubt that too many statisticians would agree with that strategy and myself would view that as something to consider but not something that should be taken as a gold standard. The fact that you have apparently not even attempted to consider two-factor interactions (I immediately would grant that you can't possibly consider all interactions). As you point out, you have substantial amounts of data. It seems that cross-validation strategies might actually be the best here,

and that you certainly could examine two-factor interactions and include relevant ones if you chose. I'm really on the fence about whether this is an absolute must, and will leave it to the editor to make a decision as to whether further modeling would be required or this deemed sufficient.

→ We have chosen 4 covariates (rheumatoid arthritis/osteoarthritis, obesity, female sex, lung cancer) and have run 4 competing risk regression analyses with interaction terms between each factor and each of the study medications (see Table 2 below). We chose these factors because they could be biologic/behavioral explanations for current HRs of tendon rupture (rheumatoid arthritis/osteoarthritis, obesity, female sex) or because the factor had an oversized HR in the current models (lung cancer). We reported not only marginal (overall) HRs of any tendon rupture but also HRs at each level of chosen covariates; e.g. HR at having obesity (Yes) and HR at not having obesity (No). Most of all interaction terms were not statistically significant except the interaction between LVX and female. While LVX exhibited 34% increased risk of tendon rupture among female (HR=1.34), it exhibited no increased risk among male. Mostly, the HRs for LEX and LVX were significantly above 1 and the HR for MXF were below 1. However, they were mostly very close to the HRs of the analysis without interactions. Inclusion of interaction terms had no influence on the overall conclusion of the paper. FQs as a class were still not associated with the increased risk of tendon ruptures. Neither CIP nor MXF exhibited any risk for tendon ruptures. LVX did exhibit significant increased risk but its risk never exceeded the risk of the non-FQ, LEX. This was a good exercise for us; offering some assurance that the inclusion of interaction terms would not change the overall conclusions. However, we would argue for not including them in the paper because they would add bulk but no additional insight to the reader.

Table 2: Comparing HRs from a model with to without interaction terms.

	Level of Interacted term	With interaction with RAOA	With interaction with Obesity	With interaction with Female	With interaction with lung cancer	No Interaction
		HR(95% CI)	HR(95% CI)	HR(95% CI)	HR(95% CI)	HR(95% CI)
AMX VS. NO	Overall	0.85(0.78,0.93)	0.86(0.80,0.92)	0.86(0.80,0.92)	0.87(0.81,0.93)	0.86(0.80,0.92)
	No	0.84(0.75,0.95)	0.87(0.81,0.94)	0.86(0.78,0.95)	0.86(0.80,0.92)	
AMX ≤	Yes	0.87(0.80,0.95)	0.80(0.68,0.95)	0.86(0.78,0.95)	1.98(0.81,4.84)	

30-days						
AMC VS. NO AMC ≤ 30-days	Overall	0.93(0.84,1.03)	0.93(0.85,1.02)	0.93(0.85,1.02)	0.93(0.85,1.02)	0.93(0.85,1.02)
	No	0.92(0.80,1.07)	0.94(0.85,1.04)	0.98(0.86,1.11)	0.93(0.85,1.02)	
	Yes	0.94(0.84,1.05)	0.89(0.72,1.09)	0.90(0.79,1.02)	0.84(0.24,2.90)	
AZM VS. NO AZM ≤ 30-days	Overall	1.00(0.93,1.08)	0.99(0.93,1.06)	0.99(0.93,1.06)	1.00(0.93,1.07)	0.99(0.93,1.06)
	No	1.02(0.92,1.13)	0.97(0.90,1.04)	0.93(0.84,1.03)	0.99(0.93,1.06)	
	Yes	0.97(0.89,1.06)	1.11(0.95,1.30)	1.03(0.95,1.12)	2.03(0.98,4.23)	
LEX VS. NO LEX ≤ 30-days	Overall	1.40(1.29,1.52)	1.32(1.22,1.42)	1.30(1.21,1.40)	1.31(1.22,1.41)	1.31(1.22,1.41)
	No	1.50(1.34,1.69)	1.33(1.23,1.45)	1.41(1.28,1.57)	1.31(1.22,1.41)	
	Yes	1.22(1.11,1.34)	1.24(1.06,1.46)	1.22(1.10,1.36)	1.20(0.36,3.98)	
LVX VS. NO LVX ≤ 30-days	Overall	1.16(1.05,1.29)	1.14(1.04,1.24)	1.15(1.05,1.26)	1.14(1.04,1.25)	1.14(1.05,1.25)
	No	1.18(1.02,1.36)	1.11(1.01,1.23)	0.93(0.81,1.08)	1.14(1.04,1.25)	
	Yes	1.13(1.01,1.26)	1.25(1.04,1.51)	1.34(1.20,1.50)	1.35(0.65,2.78)	
CIP VS. NO CIP ≤ 30-days	Overall	0.98(0.90,1.06)	0.95(0.88,1.03)	0.96(0.89,1.03)	0.96(0.89,1.03)	0.96(0.89,1.03)
	No	1.00(0.89,1.12)	0.92(0.85,1.00)	0.90(0.80,1.01)	0.96(0.89,1.03)	
	Yes	0.94(0.85,1.03)	1.11(0.95,1.29)	1.00(0.91,1.10)	1.07(0.38,3.00)	
MXF VS. NO MXF ≤ 30-days	Overall	0.54(0.31,0.94)	0.59(0.37,0.93)	0.58(0.36,0.93)	0.56(0.35,0.89)	0.59(0.37,0.93)
	No	0.49(0.22,1.07)	0.59(0.36,0.98)	0.62(0.32,1.20)	0.59(0.37,0.95)	
	Yes	0.65(0.37,1.16)	0.56(0.17,1.82)	0.56(0.29,1.07)	0.01(0.00,0.01)	

Abbreviation: RAOA: Rheumatoid Arthritis/Osteoarthritis

VERSION 4 – REVIEW

REVIEWER	Joseph Nolan Northern Kentucky University USA
REVIEW RETURNED	27-Oct-2020

GENERAL COMMENTS	If these alterations below is made, my recommendation becomes accept. There's no real reason I think this needs to come out for review again. I'm willing to accept the inclusion of unadjusted prevalence and other changes that help the reader understand the magnitude of the numbers as sufficient. I am ok with the decision not to include interaction models, but a sentence indicating the reasoning behind this would be appropriate. It does seem like the statement on page 19 "We also did not consider interactions among covariates in our analysis because of the enormous number of two way interactions (as large as 6,400) and thus the problem of overfitting. " should be updated to address what you did do.

VERSION 4 – AUTHOR RESPONSE

Reviewer(s)' Comments to Author:

Reviewer: 3

Reviewer Name: Joseph Nolan

Institution and Country:

Northern Kentucky University

USA

Competing interests: None Declared

Comments to the Author

If these alterations below is made, my recommendation becomes accept. There's no real reason I think this needs to come out for review again.

I'm willing to accept the inclusion of unadjusted prevalence and other changes that help the reader understand the magnitude of the numbers as sufficient.

I am ok with the decision not to include interaction models, but a sentence indicating the reasoning behind this would be appropriate. It does seem like the statement on page 19 "We also did not consider interactions among covariates in our analysis because of the enormous number of two way interactions (as large as 6,400) and thus the problem of overfitting. " should be updated to address what you did do.

➔ We have changed the statement on page 19 accordingly; “We also did not consider interactions among covariates in our main analysis because of the problem of overfitting. We ran 4 sensitivity analyses with interaction terms between the study medications and 4 covariates (rheumatoid arthritis/osteoarthritis, obesity, female sex, lung cancer). The inclusion of interactions did not change our conclusion of no TR risk for FQ as a class.”

VERSION 5 – REVIEW

REVIEWER	Joseph Nolan Northern Kentucky University USA
REVIEW RETURNED	31-Oct-2020

GENERAL COMMENTS	It appears that the authors made the requested minor changes. As far as I am concerned this is ready for publication.