PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	Impact evaluations of drug decriminalization and legal regulation on drug use, health and social harms: A systematic review
AUTHORS	Scheim, Ayden; Maghsoudi, Nazlee; Marshall, Zack; Churchill, Siobhan; Ziegler, Carolyn; Werb, Dan

VERSION 1 – REVIEW

REVIEWER	Ole Rogeberg Frisch Centre, Norway
REVIEW RETURNED	08-Nov-2019

GENERAL COMMENTS	This is a clear and well-written systematic review that assesses the published evidence on the effects of de jure decriminalization or legal regulation of one or more previously illegal drugs. In practise, the main focus is cannabis (91% of studies), with the bulk of the research coming from the US (95% of the studies). Overall, I found the article an excellent addition to the literature, with extensive supporting materials that is likely to be helpful to future research in the field. My main issue with the paper is that it is not completely clear what the aim of the paper is. Sometimes, it is to "characterize metrics used in impact evaluations to date" (as written in the strengths-and-limitations box and subchapter - page 14, line 10-12). Other times, the aim seems to be to synthesize the existing literature and point out some key findings it supports (e.g., the pre-analysis plan lists the review question as "What are the impacts of the implementation of drug decriminalization and regulation on drug availability, use and related harms?," similar to the stated aim in the first sentence of the methods chapter - page 5, line 19). This ambiguity is also evident in the paper, which has a section on metrics as well as on "impacts of decriminalization and legal regulation." My main concerns are with the sections trying to synthesize the literature to draw tentative conclusions. Here, the authors draw on their tabulation of results in the supporting materials, which simply counts the number of statistically significant estimates, with the statistically significant estimates further subdivided by whether they indicate beneficial, negative, or mixed consequences of legislative change. When they discuss impacts, they essentially discuss whether studies on a topic tend to report statistically significant results, and if they do so, whether the results lean in one or the other direction. This approach gives no indication of whether effects matter or not (e.g., are we talking about tiny effects from a precise study that is statistically significant), how credible they are, or how

	big they are. Strong evidence that e.g., serious mental illness increased substantially would be “neutralized” by a noisy, imprecise study indicating positive effects on some single disorder. This approach also gives a misleading impression of how solid an evidence base is. Many of the papers counted are competing analyses of the same legislative changes in the same jurisdictions, but 100 studies using different methods and data to assess the effects of the same set of medical marijuana laws in the US on population use are clearly not equivalent to 100 studies of how legislative changes in 100 different nations affect use. Take, for instance, the studies on legal cannabis and opioid overdose deaths, which has a back-and-forth flavour as research groups add new years of data and new policy variables that shift conclusions. As the authors note on page 14 notes (46-51), Shover et al found that the beneficial association reversed direction in later years, but Powell 2018 noted the same thing and argued that it was explained by the lack of active dispensaries. Data from earlier studies will be included in later studies, some of the models estimated will encompass earlier models as special cases, and simply counting the studies and tabulating their results does not seem like an appropriate way to summarize the literature. I could not find any discussion of bias in either individual studies or the literature as a whole, although the Prisma form indicates that this should be done on page 7 and in supplementary table 1. Is the bias assessment the same as the quality scores? The bias/quality issue would also seem relevant to the sections of the paper that discuss what the literature indicates on different issues (“impacts of...” on page 12). While not a quantitative meta-analysis, the discussion of how studies are distributed across beneficial, harmful, mixed effects and null effect is used as an indication of what the literature suggests. Given this, I would be interested in seeing whether a) positive, negative, mixed and null results (across topics) came from studies with different average quality, whether b) significant results (irrespective of sign) are associated with poorer quality studies that may have more researcher degrees of freedom, and c) how average quality differed across topics. A less important comment: The authors write that “Driving while under the influence of cannabis was most often found to increase following decriminalization or legal regulation” (page 12). I am curious as to whether this would be better phrased as “Driving with detectable THC concentrations in blood …” . Studies assessing the prevalence of detectable THC in blood are not equivalent to studies of impairment, as THC can remain detectable in blood long after impairment, especially for regular users (ref Huestis <https://link.springer.com/chapter/10.1007/3-540-26573-2_23>). Impairment as measured by psychomotoric tests, on the other hand, seems to be largely gone 4-6 hours after smoking. While I agree that increased use would most likely increase cannabis impaired driving, it would increase the share of crash involved drivers with detectable THC even if it didn’t - and at least the couple of studies I looked up from the reference list seemed to be estimates of the prevalence of positive THC results more than studies with measures of impairment.

	In the discussion on page 14 with caveats (line 10 to 31), it may be worth noting that there is evidence that citizens may be unaware of the specifics of the legal scheme that applies to them (see https://gspp.berkeley.edu/assets/uploads/research/pdf/ssrn-id1120930.pdf)

REVIEWER	Silviamaria Minozzi Ruffini Department of Epidemiology, Lazio Regional Health Service, Rome Italy
REVIEW RETURNED	10-Dec-2019

GENERAL COMMENTS	The review addresses a very relevant, current and interesting topic. The review has been conducted following high methodological standards. The limitation of the review has been fully discussed (e.g. the limitation of the counting of statistical significant results)

REVIEWER	Wayne Hall Centre for Youth Substance Abuse, University of Queensland, Australia
REVIEW RETURNED	20-Jan-2020

GENERAL COMMENTS	Review for BMJ Open Drug decriminalisation This paper provides a very useful overview of empirical studies of the impacts of various types of drug liberalisation (decriminalisation, legalisation etc) on indicators of drug use and drug-related harms. The authors have adopted best practice in conducting their systematic searches of the literature and in their methods of data extraction. The methods used to rate study quality are clearly described. The review findings are perhaps not too surprising to someone familiar with the field but it is nonetheless useful to have one's impressions created by less extensive and systematic searches confirmed by a literature search that has been conducted in a systematic way and which provides a detailed summary of the type of policies that have been evaluated, the main types of study designs used, the public health outcomes that have been measured, and the quality of the studies. The main findings are that studies of cannabis liberalisation in the USA are the largest category of study type, with many more studies done on the impacts of legal cannabis regulation than on cannabis decriminalisation (despite the latter being a more common form of policy liberalisation). The dominant outcomes in these studies have been measures of lifetime and past year use of cannabis and other drug use. There are very few studies of the effects of policy changes on outcomes of public health significance, such as problem drug use, car crashes etc. The study designs are also weak in often comparing time series of aggregate outcomes (e.g. state level prevalence) with crude comparisons between states that have or have not legalised medical or adult use of cannabis.

	The discussion could briefly explain why the literature is dominated by US studies of changes in the prevalence of cannabis use that have used weak study designs and limited statistical methods. 1. A major factor is that the most radical form of cannabis liberalisation (legalisation of commercial supply for medical and non medical use) has primarily occurred in the USA and more recently in Canada. 2. There has been limited funding for drug policy research so most policy evaluations comprise opportunistic analyses of publicly available data sets. Given this, it is not surprising that most studies have come from the USA where there is an abundance of publicly available national survey data on the prevalence of cannabis and other drug use. Nor is it surprising that the study designs primarily consist of comparisons of time series data from different states with state cannabis policies crudely coded into legal or not for medical or adult use. 3. The authors' box score count of outcomes of the impacts of liberal medical cannabis laws on prevalence of cannabis use have missed the reasonably consistent finding that the frequency of use has increased among existing adult cannabis users though not among adolescents (see reviews of the evidence by Hasin and colleagues on this issue). 4. A brief comment should be included on why it is too early to assess the public health effects of the legalisation of commercial cannabis supply in USA, namely, that the policy has been incompletely implemented in a small number of US states while Federal prohibition has constrained full commercialisation (see https://www.ncbi.nlm.nih.gov/pubmed/31657733). This is important point because there is a tendency for many readers to interpret the limited evidence of legalisation's impact to date as good evidence that cannabis commercialisation will have no effects on the prevalence of problem use and harms in the longer term.

REVIEWER	Alex Stevens University of Kent UK
REVIEW RETURNED	25-Jan-2020

GENERAL COMMENTS	This review provides interesting analysis of the metrics being used in peer-reviewed evaluations of drug law reform in some countries. But it seems a bit confused over whether it aims to provide a summary of the impact of these reforms. This is stated to be one of the aims on page 5. It is also implied by the title of the article. But the statement of strengths and limitations on page 3, the text on page 14 (lines 10-14), and the concluding section focus on the aim of 'characterizing metrics'.

	The analysis of metrics is useful and worthy of publication. Previous articles (e.g. Kilmer & Pacula, 2017; Stevens, Hughes, Hulme, & Cassidy, 2019) have noted the tendency to focus evaluation on prevalence of substance use (as well as the limitations of this metric). This current article is the first to quantify this phenomenon in a systematically selected sample of studies. The aim to 'summarize the impacts of decriminalization and legal regulation' is problematic for at least three reasons. One is that the quality of reviewed studies is generally low by the usual standards of a systematic review. These studies' findings are also not reported in ways that allow cumulation of results, which leaves the authors here to use 'vote-counting'. The limitations of this are well discussed in the 'strengths and limitations' section. A second problem with the aim of summarizing impacts is that most of the evaluations relate only to the first few months and years of reforms, of which the effects may not be apparent in the short term (Dirisu, Shickle, & Elsey, 2016). As we have recently seen with the cited study by Shover et al, early results may not be a reliable indicator of the eventual impacts of reforms. A third - and most problematic - issue with summarizing impacts of decriminalization and legal regulation relates to the differences between these two types of policy reform. As defined here, decriminalization relates to offences of possession, while legal regulation relates to modes of supply. So different mechanisms are involved. There are good reasons to think that they may have different effects (see MacCoun & Reuter, 2011). The heterogeneity is acknowledged but, according to the statement of strengths and limitations 'was not considered in this review'. This is a puzzling decision, which causes some problems in the analysis. This is partly because of the preponderance of studies of legal regulation found by the search strategy. They dominate evaluations of decriminalization for every reported aspect of impact. Take, for example, the six studies that are reported to attribute increases in hospitalizations to 'decriminalization or legalization'. Five of these studies are about legalization. The findings of the one study of decriminalization are not adequately represented by the report of an attribution of increased hospitalizations; Model (1993) did report an increase for marijuana-related hospitalizations, but also a reduction for other drugs. Similarly, four of the five studies that are reported as showing an increases of driving while under the influence of cannabis relate to legal regulation. The one study of decriminalization (by Pollinit et al) actually reported 'discrepant findings regarding the impact of decriminalization on marijuana-involved driving'. The presentation of finding about legal regulation as if they also apply to decriminalization is unjustified. The disparity between the number of studies found on legal regulation compared to decriminalization (and their concentration in the USA) is partly caused by the search strategy adopted, as well as some questionable applications of this strategy. The inclusion criteria are stated as including original, peer-reviewed, observational studies which use quantitative measurements before and after implementations of decriminalization or legal regulation. But at least two such studies of Portugal (Hughes & Stevens, 2010; Laqueur, 2014) are not included in the review. Both these

	studies are peer-reviewed and they do include some before and after measurements. Indeed, their inclusion would have increased by 50% the number of studies which reported before and after measures of drug-related criminal justice involvement. It might be argued that these are not 'original' studies, as they rely on secondary data. But so do many of the studies included in the review. The absence of several countries that have decriminalised drug possession (including Italy, Spain, Germany and Estonia) from the review can be attributed to the decision not to include grey literature or book chapters. It may have been exacerbated by the approach taken to publications in languages other than English. Much of the research that has been written on the effects of decriminalizations in Europe appears in book chapters (e.g. Holzer, 2017; Quintas & Arana, 2017) or in reports (e.g. Solivetti, 2001; Zábranský, Mravčík, Gajdošíková, & Milovský, 2001). These contain valuable information which would have added to the picture given of both the range of impacts and the metrics used to capture them. There is also useful information on the 1994 decriminalization of possession of small quantities of drugs by the German federal constitutional court, but much of this is in German and was apparently not found or included in the search for this review. There are German-language peer-reviewed journals (e.g. Suchttherapie) that have published articles on decriminalization, but I do not know if these are indexed in the databases that were searched for the review. The inclusion of Belgium and France in the review does not match its definition of decriminalization as the removal of criminal penalties. In neither country were criminal penalties removed for possession. As the cited study by Adams and Raschzok reports, France just made it optional for these offences to be prosecuted, and Belgium replaced prison sentences with fines. If these are forms of decriminalization, so then would be other countries included in their analysis by Hughes, Matias and Griffiths (2018), which is also a peer-reviewed study that includes before and after measurements. Studies before and after the 1976 Dutch decision to stop prosecuting possession of small quantities of 'soft' drugs would also have to be considered. Some more minor points: - The results section of the abstract refers to 'all three substance use metrics'. But the preceding sentence refers to four such metrics. - The introduction could do more to inform readers of the existence of previous reviews in this area (e.g. Ammerman, Ryan, & Adelman, 2015; Dirisu et al., 2016; Kilmer & Pacula, 2017; Kotlaja & Carson, 2018; Pacula et al., 2005; Stevens et al., 2019; Waddell & Wilson, 2017). - It appears that the only Australia studies included relate to South Australia, which seems odd, given that ACT and the Northern Territories also have de jure forms of decriminalization which have been evaluated (Hughes et al., 2016). - The study by Mauro et al 2017 of perceived availability is incorrectly cited as an example of a study that does not include any other metric than the prevalence of use. As one of the cited studies by Pacula et al suggests, the devil in evaluating drug law reforms is in the detail. I am therefore not sure

	whether it makes sense to attempt to briefly summarise the impacts of heterogeneous reforms across two broad and different types. Overall, I believe that this review would make a valuable contribution to the literature if it were more strictly focused on the issue of metrics used in peer-reviewed evaluations of drug law reforms, rather than on summarizing impacts. An alternative would be to separate out the findings about decriminalization and legal regulation so that the latter do not inappropriately dominate the former, or even to exclude the small number of evaluations of decriminalization, which would leave more space for discussing differences in legal regulation (e.g. between medical marijuana laws and the legalization of cannabis for 'recreational' use). References Ammerman, S., Ryan, S., & Adelman, W. P. (2015). The Impact of Marijuana Policies on Youth: Clinical, Research, and Legal Update. Pediatrics, 135(3), e769–e785. Dirisu, O., Shickle, D., & Elsey, H. (2016). Influence of legal status on the uptake of cannabis in young people. Current Opinion in Psychiatry, 29(4), 231–235. Holzer, T. (2017). German drug policy. In R. Colson & H. Bergeron (Eds.), European Drug Policies: The Ways of Reform (pp. 100–113). Hughes, B., Matias, J., & Griffiths, P. (2018). Inconsistencies in the assumptions linking punitive sanctions and use of cannabis and new psychoactive substances in Europe. Addiction. Hughes, C. E., & Stevens, A. (2010). What can we learn from the Portuguese decriminalization of illicit drugs? British Journal of Criminology, 50(6), 999–1022. Hughes, C.E., Ritter, A., Chalmers, J., Lancaster, K., Barrat, M., & Moxham-Hall, V. (2016). Decriminalisation of drug use and possession in Australia – A briefing note. Sydney: Drug Policy Modelling Project, UNSW Australia. Kilmer, B., & Pacula, R. L. (2017). Understanding and learning from the diversification of cannabis supply laws. Addiction, 112(7), 1128–1135. Kotlaja, M. M., & Carson, J. V. (2018). Cannabis Prevalence and National Drug Policy in 27 Countries: An Analysis of Adolescent Substance Use. International Journal of Offender Therapy and Comparative Criminology Laqueur, H. (2014). Uses and Abuses of Drug Decriminalization in Portugal. Law & Social Inquiry. MacCoun, R. J., & Reuter, P. (2011). Assessing Drug Prohibition and Its Alternatives: A Guide for Agnostics. Annual Review of Law and Social Science, 7(1), 61–78. Pacula, R. L., MacCoun, R., Reuter, P., Chriqui, J., Kilmer, B., Harris, K., Schäfer, C. (2005). What does it mean to decriminalize marijuana? A cross-national empirical examination. Advances in Health Economics and Health Services Research, 16, 347–369. Quintas, J., & Arana, X. (2017). Decriminalization: Different Models in Portugal and Spain. In E. U. Savona, M. A. R. Kleiman, & F. Calderoni (Eds.), Dual Markets (pp. 121–143). Berlin: Springer Solivetti, L. M. (2001). Drug Use Criminalization v. Decriminalization: An Analysis in the Light of the Italian Experience. Bern: Swiss Federal Office of Public Health Stevens, A., Hughes, C.E., Hulme, S., & Cassidy, R. (2019). Depenalisation, diversion and decriminalisation: a realist review

	and programme theory of alternatives to criminalisation for simple drug possession. European Journal of Criminology Waddell, K., & Wilson, M. G. (2017). Rapid Synthesis: Examining the Impact of Decriminalizing or Legalizing Cannabis for Recreational Use. Hamilton, Canada: McMaster Health Forum/Michael G. DeGroote Centre for Medicinal Cannabis Research. Zábranský, T., Mravčík, V., Gajdošíková, H., & Milovský, M. (2001). Impact analysis project of new drugs legislation. Summary final report. Prague: Office of the Czech Government, Secretariat of the National Drug Commission.

REVIEWER	Professor Catherine Comiskey Trinity College Dublin The University of Dublin Ireland
REVIEW RETURNED	29-Jan-2020

GENERAL COMMENTS	Thank you for the opportunity to review this worthwhile paper. I think it adds to the literature and addresses a gap. However, I think for the benefit of a wider readership the article requires greater clarity and explanation in some areas. Please see my detailed comments below. Abstract line 39/40 Across all three substance use metrics, legal reform was most often statistically unassociated with use Please clarify the meaning of the above assertion. Is the reader to infer that legal reform was not associated with use, is that to say there was no association between legal reform and use yes or no or no association between legal reform and use increase or use decrease? Abstract line 43 forward The conclusion summarises the findings and recommends improvement in drug law reform evaluation but can it go further? What can the authors conclude about the association between legal change and prevalence or legal change and health? Page 3 I agree with the strengths and limitations Page 4 line 33 typing error '2' Page 5 lines, 10,11 and 12. The aim here 'We specifically aimed to characterize the topical and geographic range of existing studies, summarize the impacts of decriminalization and legal regulation, and identify gaps in the evidence' is different to that stated in the abstract. Personally, I preferred the one in the abstract but perhaps given your conclusions the aim as stated here was closer to what was done. Can you clarify please. Page 5, line 28. The protocol makes interesting reading. I found your definition of exposure useful and the information on controls also useful. Did you proceed with this approach or did you change your approach? A little more detail on the approach in the abstract would help. Page6 line 12 'Eligible studies included outcome measures pertaining to drug availability, use, or related health and social harms, following the schema developed by Nutt and colleagues.[18] ' I think readers would be interested in knowing more about the choices of measures especially if as you suggest more research needs to be done to address this gap. Page 7, lines 13 to 30. Interesting approach. In terms of categorising the changes into one of the 4 categories, (beneficial, harmful, mixed, or null). This is a key point in the approach to the analysis and given your subsequent findings I think you need to

	further justify this categorisation. Perhaps given an example as an illustration. Was the categorisation agreed upon as a valid approach by practitioners or service users. For example, who gets to decide if something is harmful? Are their harms that are lesser harms etc. I think your approach is valid, but it needs great clarity and justification given its potential importance to your later findings. Page 9, line 48. Were study outcomes included in only one of your created metrics or could they be included in more than one? Page 10, line 12, can you provide more detail on 'perceived harmfulness', what exactly is this? Page 10 lines 19 to 40. This is very interesting. From this is it true to say that from this we can see that almost no studies measured physical welling or mental wellbeing of the people who use drugs. Yet it is these very domains that form part of the key arguments for decriminalisation and legalisation. This would be a point worth exploring and is a finding in itself. Page 11, line 19 Supplementary table 2 is interesting for example the row in the table describing Prevalence or frequency of use, other drugs/alcohol. However you do not discuss the table but just say it provides tallys. Is there anything more we can learn from the table? Page 11, lines 19 to 36. These results are difficult to comprehend. Could the specifics of these particular results be placed in a table providing the reader with the details of the test carried out and the specific p values etc. This would be more informative. The term 'statistically unassociated' is not one I would see used can you simply say 'was not associated with' or something similar? Page 12, line 56 to page 13. The point, 'these findings are further confirmed by the preponderance of evidence synthesized in this review, which suggests that population prevalence of use is largely unaffected by drug policy. By contrast, drug policies may be able to influence the types of drugs that people use, drug-related risk behaviours, and modes of drug consumption.[86] Metrics to assess these outcomes, however, were lacking in the reviewed literature.' Is in my opinion very important and worth highlighting further, especially the lack of physical and mental wellbeing measurement.

REVIEWER	Brendan Hughes, Marica Ferri EMCDDA Portugal
REVIEW RETURNED	14-Feb-2020

GENERAL COMMENTS	Thank you for the opportunity to review this paper. Its findings are crucial and they map others' experience in drug law evaluations; namely (and crudely) that whatever the stated intention of the policy, which may be to save police resources or concentrate them towards organised crime, it seems the most common form of evaluation question is "did use go up?". A basic principle of evaluation is to evaluate against the initial objective of the law, but this seems too difficult, or too uninteresting, for some. EMCDDA has published a guide to drug policy evaluation which emphasises the policy model and the alignment of objectives, research questions, indicators and study types. Our main concerns are with the hypothesis, in that it is not very clear, and the technique for carrying out a systematic review.

	Our first concern is that this study looks at “decriminalisation and legal regulation”. In the study period it is known that these were not the only possible options, and some “decriminalisations” have since been reversed, or simply criminalisations have been toughened by increasing penalties. It would have been nice to hear about these also. However such inclusion would require starting the exercise from the beginning. How or why were these topics chosen? Perhaps the authors have used the recent statement on drug decriminalisation by the UN Chief Executives Board as a basis for their research question (https://www.unsceb.org/CEBPublicFiles/CEB-2018-2-SoD.pdf). We are not sure where legal regulation fits in that. Our other concern is the mixing of evaluations of policy changes on illicit drug use and medicinal drug use. Why is this done? Is it to consider any newly legal availability of a psychoactive substance – in which case certain prescription medicines could also be included, particularly noting the rise in opioid availability in the US over the time period studied? Could one also include changes in the prescription status of benzodiazepines? Or is there an insinuation (in some cases justified) that “medical cannabis” was poorly controlled and often just a cover for recreational use – noting that prescription opioids were not particularly well controlled either, with devastating consequences? This might be the case for North America but so far data shows that the medical and recreational cannabis laws in Europe are simply not comparable, prescribing for medical purposes is very tightly controlled. We feel the authors should better clarify why they have chosen to do this, or perhaps separate out the medical cannabis laws for a separate paper. The findings are valid but so far they sit rather strangely here. Otherwise, there are issues regarding the technique for carrying out a systematic review. Referring to the Methods section, P5 Lines 25-26, it should be noted that PRISMA is a quality of reporting tool, not a quality of review one: “PRISMA may also be useful for critical appraisal of published systematic reviews, although it is not a quality assessment instrument to gauge the quality of a systematic review.” http://www.prisma-statement.org/. In the Data Analysis subsection, P7 Lines 13-15, the categories of outcomes look difficult to interpret: “beneficial or harmful” for whom (individuals, societies) and under which conditions? The outcomes should have been specified before the search strategy and be considered for conceiving the search strategy, in addition they could have made measurable (like violating article X of the Convention Y). In the same way, in the Study Outcome Measures and Metrics on p9 Lines 48-50, in a Systematic Review outcomes are decided before seeing the studies; basing an analysis on existing outcomes constitutes a descriptive analysis rather than an evaluation. In this way it is difficult to define the study as a systematic review of scientific literature. However it is useful and we suggest to call it a literature review to generate hypothesis (the authors recognise some important limitations). The most important implication of the study is perhaps to ensure to have baseline data evaluation in place before the change in law is introduced, in order to prepare evaluations, in accordance with the new law’s objectives, and in turn encourage a discussion of what are the expected outcomes. Ultimately we look forward to evaluations becoming more comparable to provide a more reliable body of evidence.

VERSION 1 – AUTHOR RESPONSE

Reviewer Comment	Author Response	Page
Reviewer 1		
My main issue with the paper is that it is not completely clear what the aim of the paper is. Sometimes, it is to "characterize metrics used in impact evaluations to date" (as written in the strengths-and-limitations box and subchapter - page 14, line 10-12). Other times, the aim seems to be to synthesize the existing literature and point out some key findings it supports (e.g., the pre-analysis plan lists the review question as "What are the impacts of the implementation of drug decriminalization and regulation on drug availability, use and related harms?," similar to the stated aim in the first sentence of the methods chapter - page 5, line 19). This ambiguity is also evident in the paper, which has a section on metrics as well as on "impacts of decriminalization and legal regulation."	We have clarified that the paper has one primary and one secondary aim – first, to characterize metrics and second, to synthesize existing literature. For instance, at the end of the introduction on page 5, we now state: "Given increasing interest in quantifying the impact of drug law reform, as well as a lack of systematic assessment of outcomes beyond adolescent cannabis use to date, we conducted a systematic review of original peer-reviewed research evaluating the impacts of (a) legal regulation and (b) drug decriminalization on drug availability, use, or related health and social harms. Our primary aim is to characterize studies with respect to metrics and indicators used. The secondary aim is to summarize the findings and methodologic quality of studies to date." We recognize that the aim of characterizing metrics was not specified in the protocol; this was an oversight as we view this as a necessary first step (i.e., to describe what outcomes have been studied) before synthesizing the study findings with respect to those outcomes.	Abstract, p. 5

My main concerns are with the sections trying to synthesize the literature to draw tentative conclusions. Here, the authors draw on their tabulation of results in the supporting materials, which simply counts the number of statistically significant estimates, with the statistically significant estimates further subdivided by whether they indicate beneficial, negative, or mixed consequences of legislative change. When they discuss impacts, they essentially discuss whether studies on a topic tend to report statistically significant results, and if they do so, whether the results lean in one or the other direction. This approach gives no indication of whether effects matter or not (e.g., are we talking about tiny effects from a precise study that is statistically significant), how credible they are, or how big they are. Strong evidence that e.g., serious mental illness increased substantially would be "neutralized" by a noisy, imprecise study indicating positive effects on some single disorder. This approach also gives a misleading impression of how solid an evidence base is. Many of the papers counted are competing analyses of the same legislative changes in the same jurisdictions, but 100 studies using different methods and data to assess the effects of the same set of medical marijuana laws in the US on population use are clearly not equivalent to 100 studies of how legislative changes in 100 different nations affect use.	We thank the Reviewer for raising this important issue. We are cognizant of the limitations of simple 'vote-counting,' as we reflect in the Discussion. In response to this comment, we have made substantial revisions to increase the utility of systematically assessed data, including reporting quality scores disaggregated by metric and presenting the summaries of results from decriminalization vs. regulation studies separately. In the context of a first article from a large systematic review, we believe that a focus on all outcomes rather than a more homogenous subset is of high value. A highly granular discussion of various results is, we agree with the Reviewer, needed; however, we believe that this would be better served as a separate manuscript that builds on this initial overview of the literature, and which can accommodate meta-analyses of sub-categories of studies included in this comprehensive review. In any case, we appreciate the additional limitations highlighted here and have added them to the Discussion (new text is **bolded**):	p. 16-17

Take, for instance, the studies on legal cannabis and opioid overdose deaths, which has a back-and-forth flavour as research groups add new years of data and new policy variables that shift conclusions. As the authors note on page 14 notes (46-51), Shover et al found that the beneficial association reversed direction in later years, but Powell 2018 noted the same thing and argued that it was explained by the lack of active dispensaries. Data from earlier studies will be included in later studies, some of the models estimated will encompass earlier models as special cases, and simply counting the studies and tabulating their results does not seem like an appropriate way to summarize the literature.	"Our use of vote-counting in this synthesis (i.e., categorizing individual outcome measures as indicating beneficial, harmful, mixed/subgroup-specific, or no statistically significant associations) is subject to the same limitation. Vote-counting should also be interpreted with caution in light of the heterogeneity of outcome definitions, the inherent arbitrariness of statistical significance thresholds, and **the key distinction between statistical and clinical significance. In addition, many included studies are evaluating the same policies (e.g., cannabis legalization in western U.S. states), sometimes using overlapping data but drawing different conclusions based on analytic choices and timeframes. The existence of multiple datapoints for a particular outcome does not imply that the outcome has been well-studied across diverse contexts such that scientific consensus on its effects has been reached.** Moreover, as illustrated by a recently published extension of the included article by Bachhuber et al.,[77] multiple high-quality studies may generate results that are later revealed to be spurious as additional follow-up data become availability. Specifically, Shover et al. demonstrated that the positive association reported between medical cannabis legalization and opioid overdose mortality in 1999-2010 reversed direction in later years, suggesting that earlier findings of a protective effect should not be given causal interpretations.[96] **This was foreshadowed in the included article by Powell et al., which found that the purportedly positive effect of medical cannabis legalization was attenuated in 2010-2013.[81] This scientific back-and-forth can be expected given that most included articles are evaluating legal changes introduced rather recently, and thus are examining early impacts with limited years of follow-up."**	
I could not find any discussion of bias in either individual studies or the literature as a whole, although the Prisma form indicates that this should be done on page 7 and in supplementary table 1. Is the bias assessment the same as the quality scores?	We have clarified in the text that the Downs and Black checklist scores reflect assessment of internal validity (bias), as well as reporting and external validity:	p. 8

	"Downs and Black checklist [25] for observational studies (see Appendix B), **which assesses internal validity (bias), external validity, and reporting**."	
The bias/quality issue would also seem relevant to the sections of the paper that discuss what the literature indicates on different issues ("impacts of..." on page 12). While not a quantitative meta-analysis, the discussion of how studies are distributed across beneficial, harmful, mixed effects and null effect is used as an indication of what the literature suggests. Given this, I would be interested in seeing whether a) positive, negative, mixed and null results (across topics) came from studies with different average quality, whether b) significant results (irrespective of sign) are associated with poorer quality studies that may have more researcher degrees of freedom, and c) how average quality differed across topics.	Thank you for this great suggestion. We have now added (1) a breakdown of average quality scores by outcome metric to Supplementary Table 2; and (2) a discussion of average quality by direction of effect to the Study Quality section: "Study quality differed significantly ($p<0.001$) by the direction of the association with the outcome of interest, with higher quality scores among studies estimating mixed (X=15.4) or beneficial (X=15.2) versus null (X=14.2) or harmful (X=13.1) effects of legal change on the outcome of interest." To aid interpretation, we have also noted in the "impacts of" section where metrics came from studies with particularly low (e.g., driving after use) or high (e.g., prescription drug use) average quality scores.	Suppl Table 2, p.10, p. 12

A less important comment: The authors write that “Driving while under the influence of cannabis was most often found to increase following decriminalization or legal regulation” (page 12). I am curious as to whether this would be better phrased as “Driving with detectable THC concentrations in blood …” . Studies assessing the prevalence of detectable THC in blood are not equivalent to studies of impairment, as THC can remain detectable in blood long after impairment, especially for regular users. Impairment as measured by psychomotoric tests, on the other hand, seems to be largely gone 4-6 hours after smoking. While I agree that increased use would most likely increase cannabis impaired driving, it would increase the share of crash involved drivers with detectable THC even if it didn’t - and at least the couple of studies I looked up from the reference list seemed to be estimates of the prevalence of positive THC results more than studies with measures of impairment.	We agree with the Reviewer that this phenomenon would be better phrased as “driving with detectable concentrations”; however, one included study (Hasin et al. 2017) did rely on self-report of impaired driving. Therefore, we have renamed the outcome “driving under the influence or with detectable concentrations of the decriminalized or regulated drug”) in the paper (and “driving after use” in Figure 3 for brevity). We have also clarified in the text (p. 11) that one study used self-reported impaired driving.	Pg. 11, Figure 3,
In the discussion on page 14 with caveats (line 10 to 31), it may be worth noting that there is evidence that citizens may be unaware of the specifics of the legal scheme that applies to them.	We have added this: “Further, the impact of drug laws on drug use and related outcomes may be limited by a lack of public awareness of the details of local laws.[92]”	p. 16
Reviewer 3		

The discussion could briefly explain why the literature is dominated by US studies of changes in the prevalence of cannabis use that have used weak study designs and limited statistical methods: 1. A major factor is that the most radical form of cannabis liberalisation (legalisation of commercial supply for medical and non medical use) has primarily occurred in the USA and more recently in Canada. 2. There has been limited funding for drug policy research so most policy evaluations comprise opportunistic analyses of publicly available data sets. Given this, it is not surprising that most studies have come from the USA where there is an abundance of publicly available national survey data on the prevalence of cannabis and other drug use. Nor is it surprising that the study designs primarily consist of comparisons of time series data from different states with state cannabis policies crudely coded into legal or not for medical or adult use.	We that the Reviewer for this suggestion. At present, the first point is captured in the first paragraph of the Discussion ("… This rapid growth in scholarship was driven by the implementation and subsequent evaluation…."). However, we now address the second point in (1) the discussion of study limitations: "The use of these study designs may be related to limited resources for prospective drug policy evaluations, with many studies relying on publicly available, routinely collected data."	p. 15, p. 17

	and (2) the conclusion paragraph: “Funding will also be required to support rigorous prospective evaluations of legal reforms.”	
The authors’ box score count of outcomes of the impacts of liberal medical cannabis laws on prevalence of cannabis use have missed the reasonably consistent finding that the frequency of use has increased among existing adult cannabis users though not among adolescents (see reviews of the evidence by Hasin and colleagues on this issue).	We agree with the Reviewer on this point, though have not stratified results by age group in this manuscript as a result of heterogeneity of metrics used and effects reported. Nevertheless, we agree that this is an important point, and have expanded in the Discussion as follows: “These findings are supported by the preponderance of evidence synthesized in this review, although some variation is evident in relation to the specific provisions of legal reforms (e.g., liberal versus tightly regulated medical markets [90]). Impacts of legal cannabis regulation on prevalence and frequency of use continue to be evaluated, with recent data suggesting small increases among adults, but not youth.[91] “	p. 14

A brief comment should be included on why it is too early to assess the public health effects of the legalisation of commercial cannabis supply in USA, namely, that the policy has been incompletely implemented in a small number of US states while Federal prohibition has constrained full commercialisation This is important point because there is a tendency for many readers to interpret the limited evidence of legalisation's impact to date as good evidence that cannabis commercialization will have no effects on the prevalence of problem use and harms in the longer term.	We have added the following to the Discussion (final line): "Longer-term impacts of non-medical cannabis legalization, and how they might be influenced by increased commercialization are yet to be seen.[97]"	p. 17
Reviewer 4		
This review provides interesting analysis of the metrics being used in peer-reviewed evaluations of drug law reform in some countries. But it seems a bit confused over whether it aims to provide a summary of the impact of these reforms. This is stated to be one of the aims on page 5. It is also implied by the title of the article. But the statement of strengths and limitations on page 3, the text on page 14 (lines 10-14), and the concluding section focus on the aim of 'characterizing metrics'.	As described in our response to Reviewer 1, we have clarified the primary and secondary aims of the paper.	Abstract, p. 5

The analysis of metrics is useful and worthy of publication. Previous articles (e.g. Kilmer & Pacula, 2017; Stevens, Hughes, Hulme, & Cassidy, 2019) have noted the tendency to focus evaluation on prevalence of substance use (as well as the limitations of this metric). This current article is the first to quantify this phenomenon in a systematically selected sample of studies.		
The aim to 'summarize the impacts of decriminalization and legal regulation' is problematic for at least three reasons. One is that the quality of reviewed studies is generally low by the usual standards of a systematic review. These studies' findings are also not reported in ways that allow cumulation of results, which leaves the authors here to use 'vote-counting'. The limitations of this are well discussed in the 'strengths and limitations' section.	As described in our response to Reviewer 1, we have expanded our reporting and discussion of study quality issues.	Suppl Table 2, p.10, p. 12

A second problem with the aim of summarizing impacts is that most of the evaluations relate only to the first few months and years of reforms, of which the effects may not be apparent in the short term (Dirisu, Shickle, & Elsey, 2016). As we have recently seen with the cited study by Shover et al, early results may not be a reliable indicator of the eventual impacts of reforms.	As described in our response to Reviewers 1 and 3, we have added text to the Discussion on the preliminary nature of findings to date and the need to examine longer-term outcomes.	p. 16-17
A third - and most problematic - issue with summarizing impacts of decriminalization and legal regulation relates to the differences between these two types of policy reform. As defined here, decriminalization relates to offences of possession, while legal regulation relates to modes of supply. So different mechanisms are involved. There are good reasons to think that they may have different effects (see MacCoun & Reuter, 2011). The heterogeneity is acknowledged but, according to the statement of strengths and limitations 'was not considered in this review'. This is a puzzling decision, which causes some problems in the analysis.	We agree that there are important differences both within and between legal regulation and decriminalization policies. We have added a section to the paper that narratively summarizes findings separately for the 19 included papers that evaluated decriminalization. We have also revised the strengths and limitations section to clarify that we do not address heterogeneity *within* legal regulation or decriminalization policies.	p.3, p. 12-13

This is partly because of the preponderance of studies of legal regulation found by the search strategy. They dominate evaluations of decriminalization for every reported aspect of impact. Take, for example, the six studies that are reported to attribute increases in hospitalizations to 'decriminalization or legalization'. Five of these studies are about legalization. The findings of the one study of decriminalization are not adequately represented by the report of an attribution of increased hospitalizations; Model (1993) did report an increase for marijuana-related hospitalizations, but also a reduction for other drugs. Similarly, four of the five studies that are reported as showing an increases of driving while under the influence of cannabis relate to legal regulation. The one study of decriminalization (by Pollinit et al) actually reported 'discrepant findings regarding the impact of decriminalization on marijuana-involved driving'. The presentation of finding about legal regulation as if they also apply to decriminalization is unjustified.	As above, the details of decriminalization studies (including the findings of Model 1993) are now discussed in the Results.	p. 12-13
The disparity between the number of studies found on legal regulation compared to decriminalization (and their concentration in the USA) is partly caused by the search strategy adopted, as well as some questionable applications of this strategy. The inclusion criteria are stated as including original, peer-reviewed, observational studies which use quantitative measurements before and after implementations of decriminalization or legal regulation. But at least two such studies of Portugal (Hughes & Stevens, 2010; Laqueur, 2014) are not included in the review. Both these studies are peer-reviewed and they do include some before and after measurements. Indeed, their inclusion would have increased by 50% the number of studies which reported before and after measures of drug-related criminal justice	With regards to the search strategy, comprehensive search terms for both legal regulation and decriminalization were used. The medical subject headings (MeSH) subheadings "Legislation & Jurisprudence" and "Supply & Distribution" were applied, along with the following keywords: legal*, decriminal*, depenaliz*, depenalis*, deregulat*, liberaliz*, liberalis*, policy, policies, or law, laws, licens*, legislation, or regulat*.	p. 6, p. 15

involvement. It might be argued that these are not 'original' studies, as they rely on secondary data. But so do many of the studies included in the review.	Both Hughes & Stevens, 2010 and Laqueur, 2014 were included in our search results and screened for inclusion but unanimously excluded for not being original research. We did not consider as original research studies that reproduced secondary data without conducting original quantitative analyses of the data. We have clarified this in the Methods, and also acknowledge in the Limitations that our decision to focus on peer-reviewed, original research did limit the geographic scope of included studies. We nevertheless maintain that a focus on such research is warranted given the weight given to it in evidence-based policymaking and its greater accessibility.	

The absence of several countries that have decriminalised drug possession (including Italy, Spain, Germany and Estonia) from the review can be attributed to the decision not to include grey literature or book chapters. It may have been exacerbated by the approach taken to publications in languages other than English. Much of the research that has been written on the effects of decriminalizations in Europe appears in book chapters (e.g. Holzer, 2017; Quintas & Arana, 2017) or in reports (e.g. Solivetti, 2001; Zábranský, Mravčík, Gajdošíková, & Milovský, 2001). These contain valuable information which would have added to the picture given of both the range of impacts and the metrics used to capture them. There is also useful information on the 1994 decriminalization of possession of small quantities of drugs by the German federal constitutional court, but much of this is in German and was apparently not found or included in the search for this review. There are German-language peer-reviewed journals (e.g. Suchttherapie) that have published articles on decriminalization, but I do not know if these are indexed in the databases that were searched for the review.	Systematic reviews of the scientific literature are usually limited to academic journals, and while we acknowledge that this limits the geographic scope of included publications, we think this limitation is reasonable as described above. The point about indexing is an important one, and we have added that to the limitations (added text in bold): "All eligible articles identified were published in English; this may reflect a paucity of evaluation research published in other languages and/or limitations of our search strategy (e.g., **some non-English journals may not be indexed in the 10 databases searched**)." We have also added to the Discussion a recommendation for scoping reviews that would include a greater diversity of evaluations:	pg. 15

	"Scoping reviews inclusive of grey literature would be valuable for describing the full range of evaluations that have been conducted globally."	
The inclusion of Belgium and France in the review does not match its definition of decriminalization as the removal of criminal penalties. In neither country were criminal penalties removed for possession. As the cited study by Adams and Raschzok reports, France just made it optional for these offences to be prosecuted, and Belgium replaced prison sentences with fines. If these are forms of decriminalization, so then would be other countries included in their analysis by Hughes, Matias and Griffiths (2018), which is also a peer-reviewed study that includes before and after measurements. Studies before and after the 1976 Dutch decision to stop prosecuting possession of small quantities of 'soft' drugs would also have to be considered.	As described on page 6, "We defined decriminalization as the removal of criminal penalties for drug use and/or possession (allowing for civil or administrative sanctions)". Following this definition, Belgium qualifies while France does not. This aligns with classifications employed by EMCDDA and the coding we used for Figure 2. We thank the Reviewer for raising this issue and, in response, have revised to remove France from the list of included countries.	p. 8, p. 11, Table 1
The results section of the abstract refers to 'all three substance use metrics'. But the preceding sentence refers to four such metrics.	We were not counting perceived harmfulness as a "substance use metric" as it is an attitudinal measure only. For clarity we have rephrased this to read "Across all substance use metrics…"	Abstract

The introduction could do more to inform readers of the existence of previous reviews in this area (e.g. Ammerman, Ryan, & Adelman, 2015; Dirisu et al., 2016; Kilmer & Pacula, 2017; Kotlaja & Carson, 2018; Pacula et al., 2005; Stevens et al., 2019; Waddell & Wilson, 2017).	Thanks for this great suggestion. We have reviewed all the suggested additions and opted to incorporate those that are peer-reviewed reviews on outcomes of decriminalization or legalization policies. We therefore added the text below to the Introduction (new text is **bolded**). "**Reviews in the scientific literature, including two meta-analyses, are narrowly focused on adolescent cannabis use. Dirisu et al. found no conclusive evidence that cannabis legalization for medical or recreational purposes increases cannabis use by young people.[20] In the two meta-analyses,** Sarvet et al…."	p. 4
It appears that the only Australia studies included relate to South Australia, which seems odd, given that ACT and the Northern Territories also have de jure forms of decriminalization which have been evaluated (Hughes et al., 2016).	The included article by Williams and Bretteville-Jensen (2014) evaluated decriminalization in South Australia, the ACT, Northern Territories, and Western Australia. Our search did not reveal any other studies from ACT or the Northern Territories meeting inclusion criteria. Williams, J., & Bretteville-Jensen, A. L. (2014). Does liberalizing cannabis laws increase cannabis use? *Journal of Health Economics*, *36*, 20–32.	

The study by Mauro et al 2017 of perceived availability is incorrectly cited as an example of a study that does not include any other metric than the prevalence of use.	Thank you for noting this error, which we have corrected (bringing the total number of outcomes to 224, as perceived availability should have been extracted separately from this study).	p. 10, p. 12
As one of the cited studies by Pacula et al suggests, the devil in evaluating drug law reforms is in the detail. I am therefore not sure whether it makes sense to attempt to briefly summarise the impacts of heterogeneous reforms across two broad and different types. Overall, I believe that this review would make a valuable contribution to the literature if it were more strictly focused on the issue of metrics used in peer-reviewed evaluations of drug law reforms, rather than on summarizing impacts. An alternative would be to separate out the findings about decriminalization and legal regulation so that the latter do not inappropriately dominate the former, or even to exclude the small number of evaluations of decriminalization, which would leave more space for discussing differences in legal regulation (e.g. between medical marijuana laws and the legalization of cannabis for 'recreational' use).	As mentioned above, we now summarize the results of decriminalization studies separately.	p. 12-13
Reviewer 5		
Abstract line 39/40 Across all three substance use metrics, legal reform was most often statistically unassociated with use	We have clarified (new text in bold):	Abstract

Please clarify the meaning of the above assertion. Is the reader to infer that legal reform was not associated with use, is that to say there was no association between legal reform and use yes or no or no association between legal reform and use increase or use decrease?	"Across all substance use metrics, legal reform was most often **not associated with changes in use**."	
Abstract line 43 forward The conclusion summarises the findings and recommends improvement in drug law reform evaluation but can it go further? What can the authors conclude about the association between legal change and prevalence or legal change and health?	While we appreciate this suggestion, word count limits for the abstract preclude us from expanding upon these important points.	
Page 5 lines, 10,11 and 12. The aim here 'We specifically aimed to characterize the topical and geographic range of existing studies, summarize the impacts of decriminalization and legal regulation, and identify gaps in the evidence' is different to that stated in the abstract. Personally, I preferred the one in the abstract but perhaps given your conclusions the aim as stated here was closer to what was done. Can you clarify please.	As described in our response to Reviewer 1, we have clarified the primary and secondary aims of the paper.	Abstract, pg. 5

Page 5, line 28. The protocol makes interesting reading. I found your definition of exposure useful and the information on controls also useful. Did you proceed with this approach or did you change your approach? A little more detail on the approach in the abstract would help.	Thank you. We have included the exposure definition (paraphrased from the protocol) on pg. 6. We had planned to stratify results by the drug being decriminalized or regulated but ultimately this did not make sense considering how few articles examined drugs other than cannabis.	
Page6 line 12 'Eligible studies included outcome measures pertaining to drug availability, use, or related health and social harms, following the schema developed by Nutt and colleagues.[18] ' I think readers would be interested in knowing more about the choices of measures especially if as you suggest more research needs to be done to address this gap.	We thank the Reviewer for this helpful suggestion. We have expanded as follows: "We used the schema developed by Nutt and colleagues to conceptualize health and social harms, including those to users (physical, psychological, and social) and to others (injury or social harm).[18]"	p. 6
Page 7, lines 13 to 30. Interesting approach. In terms of categorising the changes into one of the 4 categories, (beneficial, harmful, mixed, or null). This is a key point in the approach to the analysis and given your subsequent findings I think you need to further justify this categorisation. Perhaps given an example as an illustration. Was the	Thank you for this important comment. We agree that additional explanation would be helpful. We have added an explanation of how we coded outcomes that were neither clearly positive nor negative:	p. 7

categorisation agreed upon as a valid approach by practitioners or service users. For example, who gets to decide if something is harmful? Are their harms that are lesser harms etc. I think your approach is valid, but it needs great clarity and justification given its potential importance to your later findings.	"Although any use of cannabis and other psychoactive drugs need not be problematic at the individual level, we categorized drug use as a negative outcome given that population-level increases in use may correspond to increases in negative consequences; we thought that this cautious approach to categorization was appropriate given that such increases are generally conceptualized as negative within the scientific literature. For outcomes that are not unambiguously negative or positive, the coding approach was pre-determined taking a societal perspective. For example, increased healthcare utilization (e.g., hospital visits due to cannabis use) was coded as negative because of the increased burden placed on healthcare systems."	
Page 9, line 48. Were study outcomes included in only one of your created metrics or could they be included in more than one?	Each measured outcome was included in only one metric; we now clarify this in the methods: "We coded each study-level outcome measure into one metric grouping…"	p. 7
Page 10, line 12, can you provide more detail on 'perceived harmfulness', what exactly is this?	This refers to the perceived harmfulness to one's health of using the drug, which we have clarified.	p.10
Page 10 lines 19 to 40. This is very interesting. From this is it true to say that from this we can see that almost no studies measured physical welling or mental wellbeing of the people who use drugs. Yet it is these very	Thanks for highlighting this. We have added the following in the Discussion:	p.15

domains that form part of the key arguments for decriminalisation and legalisation. This would be a point worth exploring and is a finding in itself.	"Similarly, improving the physical and mental health of people who (already) use drugs is a motivation for drug policy reform but no included studies examined mental or physical health outcomes (aside from substance use disorders) in this population."	
Page 11, line 19 Supplementary table 2 is interesting for example the row in the table describing Prevalence or frequency of use, other drugs/alcohol. However you do not discuss the table but just say it provides tallys. Is there anything more we can learn from the table?	In the section titled "Impacts of Decriminalization and Legal Regulation" we summarize findings in Suppl Table 2. To make this clearer, the opening sentence has been revised to read: "Supplementary Table 2 tallies findings and average quality scores for each of the metrics; here we summarize findings for metrics examined in more than 5% of studies, in descending order based on the number of datapoints."	p.11
Page 11, lines 19 to 36. These results are difficult to comprehend. Could the specifics of these particular results be placed in a table providing the reader with the details of the test carried out and the specific p values etc. This would be more informative. The term 'statistically unassociated' is not one I would see used can you simply say 'was not associated with' or something similar?	We have simplified "statistically unassociated" as "not associated". Results of individual studies, including statistical results, are provided in Supplementary Table 1 (we now remind readers of this at the beginning of the section). We are happy to make additional changes to improve comprehensibility if any details in the text remain unclear.	p. 11

Page 12, line 56 to page 13. The point, 'these findings are further confirmed by the preponderance of evidence synthesized in this review, which suggests that population prevalence of use is largely unaffected by drug policy. By contrast, drug policies may be able to influence the types of drugs that people use, drug-related risk behaviours, and modes of drug consumption.[86] Metrics to assess these outcomes, however, were lacking in the reviewed literature.' Is in my opinion very important and worth highlighting further, especially the lack of physical and mental wellbeing measurement.	We have further emphasized this gap in the following paragraph, as noted above.	p. 15
Reviewer 6		
Our first concern is that this study looks at "decriminalisation and legal regulation". In the study period it is known that these were not the only possible options, and some "decriminalisations" have since been reversed, or simply criminalisations have been toughened by increasing penalties. It would have been nice to hear about these also. However such inclusion would require starting the exercise from the beginning. How or why were these topics chosen? Perhaps the authors have used the recent statement on drug decriminalisation by the UN Chief Executives Board as a basis for their research question. We are not sure where legal regulation fits in that.	Our research question was informed by the recent proliferation of research specifically on drug law liberalization and the increasing number of jurisdictions implementing decriminalization or legal regulation. While we agree that re-criminalization or enhanced criminalization policies are also worthy of study, we do see a global trend towards liberalization (as reflected in the UN CEB statement) and were specifically interested in examining the evidence base related to such policies.	p. 4

	However, we note that the recent statements on drug decriminalization by the UN Chief Executives Board (as well as position statements by the UNODC) provides important context to this review, and we therefore now refer to it in the introduction section: "Given ongoing interest by states in drug law reform, **as well as** the **recent position statement by the UN Chief Executives Board endorsing drug decriminalization,** a comprehensive understanding of their impacts to date is required.	
Our other concern is the mixing of evaluations of policy changes on illicit drug use and medicinal drug use. Why is this done? Is it to consider any newly legal availability of a psychoactive substance – in which case certain prescription medicines could also be included, particularly noting the rise in opioid availability in the US over the time period studied? Could one also include changes in the prescription status of benzodiazepines? Or is there an insinuation (in some cases justified) that "medical cannabis" was poorly controlled and often just a cover for recreational use – noting that prescription opioids were not particularly well controlled either, with devastating consequences? This might be the case for North America but so far data shows that the	We thank the Reviewer for this thoughtful comment. We sought to consider policies that would make formerly illegal classes of drugs accessible without criminal penalty or through a legal regime; we think this differs from making already legal classes of drugs more widely available and is aligned with accepted definitions of decriminalization and legalization. We added mention of the "formerly illegal" criterion in the Methods.	p. 6, p. 14

medical and recreational cannabis laws in Europe are simply not comparable, prescribing for medical purposes is very tightly controlled. We feel the authors should better clarify why they have chosen to do this, or perhaps separate out the medical cannabis laws for a separate paper. The findings are valid but so far they sit rather strangely here.	As the Reviewer notes, it is also the case that medical and recreational cannabis legalization in the U.S. cannot be so neatly disaggregated. Indeed, that isn't the situation elsewhere, and had we ultimately included any articles evaluating medical cannabis legalization outside the U.S., we would have addressed this more directly. It is possible that we did not locate such studies precisely because the tightly controlled regimes limit potential population-level effects. In response to the Reviewer's comment, we have now also added the following text to the discussion: "Importantly, the lack of non-U.S. studies evaluating legal regulation of cannabis for medical use may reflect the more tightly controlled nature of medical cannabis regulation in other countries, and thus the more limited potential for population-level effects."	

Otherwise, there are issues regarding the technique for carrying out a systematic review. Referring to the Methods section, P5 Lines 25-26, it should be noted that PRISMA is a quality of reporting tool, not a quality of review one: “PRISMA may also be useful for critical appraisal of published systematic reviews, although it is not a quality assessment instrument to gauge the quality of a systematic review.”	We appreciate the comment and have now clarified that PRISMA guidelines were followed in “preparing this manuscript”.	p. 5
In the Data Analysis subsection, P7 Lines 13-15, the categories of outcomes look difficult to interpret: “beneficial or harmful” for whom (individuals, societies) and under which conditions? The outcomes should have been specified before the search strategy and be considered for conceiving the search strategy, in addition they could have made measurable (like violating article X of the Convention Y).	As described in our response to Reviewer 5, we have expanded on the criteria used to classify outcomes as beneficial, harmful, etc., by clarifying that a societal perspective was taken and providing examples of how we coded outcomes that may not be so clearly positive or negative. No outcomes were specified in the search strategy, based on the advice of a library sciences expert on our team (Ziegler), given the wide range of potential outcomes of interest. To make this clear, we have now explicitly stated the following in the Methods: “Search terms specific to health and social outcomes were not employed so that the search would capture the broad range of outcomes of interest.”	p. 5, p. 7

In the same way, in the Study Outcome Measures and Metrics on p9 Lines 48-50, in a Systematic Review outcomes are decided before seeing the studies; basing an analysis on existing outcomes constitutes a descriptive analysis rather than an evaluation.	We thank the Reviewer for this comment. We note that, as described in the Methods (p. 7), we pre-specified 24 outcome categories; new categories were only added iteratively where outcome measures did not fit within *a priori* specified categories. Considering that the review was explicitly interested in evaluations of any health and social outcomes, we did not consider it appropriate to exclude articles because they did not fit a pre-specified category.	
In this way it is difficult to define the study as a systematic review of scientific literature. However it is useful and we suggest to call it a literature review to generate hypothesis (the authors recognise some important limitations).	As noted in our response to Reviewer 1, in the context of this first article from our systematic review that focuses on all outcomes rather than a more homogenous subset, we do think that this high-level summary of the findings has value, and we intend to publish meta-analyses of sub-categories of studies reporting on more homogeneous outcomes in the future. Nevertheless, we have made revisions to increase the utility of this data, including reporting quality scores disaggregated by metric and summarizing results from decriminalization studies separately.	

The most important implication of the study is perhaps to ensure to have baseline data evaluation in place before the change in law is introduced, in order to prepare evaluations, in accordance with the new law's objectives, and in turn encourage a discussion of what are the expected outcomes.	We wholeheartedly agree with this comment. We think the recommendation to plan for evaluation before the law is introduced is an important one and we have therefore added a sentence to the conclusion that addresses this (in response to Reviewer 4).	p. 17

VERSION 2 – REVIEW

REVIEWER	Ole Rogeberg The Ragnar Frisch Centre for Economic Research
REVIEW RETURNED	24-Jun-2020

GENERAL COMMENTS	The authors have done a good job in addressing my concerns from the first review round.

REVIEWER	Alex Stevems University of Kent
REVIEW RETURNED	06-Jun-2020

GENERAL COMMENTS	This is a revised version of an article which I previously reviewed. In my review of that submission I raised some concerns that, in my view, needed to be addressed before publication. Having read the revised version, it is now my view that it goes some way to addressing these concerns, but not far enough to warrant publication. Specifically: - the article still includes a confusing statement as to what its aims are. The introduction now includes the statement that 'Our primary aim is to characterize studies with respect to metrics and indicators used. The secondary aim is to summarize the findings and methodologic quality of studies to date'. But the discussion section states, 'This narrative synthesis did not focus on estimating the outcome-specific effects of particular decriminalization or legal regulation policies but instead sought to characterize the metrics employed to date.' If the study had primary and secondary aims, then the conclusion should clearly report its primary and secondary findings. - The article still conflates the results of studies of legal regulation and decriminilization. The authors have added a section on decriminalization. But they have retained the section on 'impacts of decriminalization and legal regulation' in which the reported effects of legal regulation dominate those of decriminlization. The discrepant findings of the studies by Model et al on hospitalizations and Pollini et al on driving are still not adequately reported.

	- the article still makes a rather arbitrary distinction between studies that are included or excluded on the basis of whether they are 'original'. It state, 'We did not consider as original research studies that reproduced secondary data without conducting original statistical analyses of the data.' As many of the included studies used secondary data, this leaves the application of statistical analysis as the definition of whether research counted as 'original'. But such statistical analysis is not always appropriate as a method for comparative, longitudinal studies of the effects of drug policy changes, as it requires a relatively large sample of cases and comparators. For example, the study by Hughes et al. of 'Inconsistencies in the assumptions linking punitive sanctions and use of cannabis and new psychoactive substances in Europe' includes a sample of only 7 countries. It would not have been appropriate to apply inferential statistical tests to this sample, yet the study provide valuable information on the effects of legal regulations on levels of drug use (or, rather, the lack of such an effect). Given that the USA is the only country in the world which systematically publishes data on drug prevalence across a reasonably large sample of geographic units, it is unsurprising that this definition of 'original' research has excluded much research from other countries. There are studies that use statistical analysis of data from large cross-national surveys (e.g. ESPAD, HBSC) on the effects of national policy regimes on drug use prevalence, but these have not been included in this review due to their cross-sectional design. - the article still does not acknowledge the contribution or existence of some previous reviews on the topic. It uses the reviews by Dirisu, Sarvet, Melchior and their co-authors to make the point about the narrow focus of previous reviews on adolescent cannabis use. However, some of the other reviews (as cited in my previous review of this article) have a broader focus than this. - France has (correctly) been removed from the list of countries that have adopted decriminalization. However, Belgium is still included, despite still using criminal penalties for possession.

REVIEWER	Catherine Comiskey Trinity College Dublin, The University of Dublin, Ireland
REVIEW RETURNED	05-Jun-2020

GENERAL COMMENTS	Thank you for your revisions. The manuscript is improved and I look forward to its publication. I am sure it will be of interest to readers and it also highlights that we need to do more research of the impact of drug use decriminalisation and legalisation on the health and well-being of people who use drugs.

VERSION 2 – AUTHOR RESPONSE

Reviewer 1

The article still includes a confusing statement as to what its aims are. The introduction now includes the statement that 'Our primary aim is to characterize studies with respect to metrics and indicators used. The secondary aim is to summarize the findings and methodologic quality of studies to date'. But the discussion section states, 'This narrative synthesis did not focus on estimating the outcome-specific effects of particular decriminalization or legal regulation policies but instead sought to characterize the metrics employed to date.' If the study had primary and secondary aims, then the conclusion should clearly report its primary and secondary findings.	We concur that the sentence in question is confusing and contradicts our statement of aims; we have therefore deleted it. As we are already over the suggested word count for BMJ Open papers we have not expanded coverage of the secondary objective in the Discussion. Of course, we are happy to do so if the Editor wishes and if additional word count can be allowed.	Pg 15, para 2

The article still conflates the results of studies of legal regulation and decriminalization. The authors have added a section on decriminalization. But they have retained the section on 'impacts of decriminalization and legal regulation' in which the reported effects of legal regulation dominate those of decriminalization. The discrepant findings of the studies by Model et al on hospitalizations and Pollini et al on driving are still not adequately reported.	Intending to balance the needs for a high-level aggregate synthesis of findings and to avoid drowning out the minority of studies from non-US locales and of decriminalization, we structured the results geographically (i.e., results for all studies were followed by results of non-US studies to ensure adequate attention was paid to these settings) and for type of legal change (i.e., all studies were followed by studies assessing decriminalization only, again to ensure that adequate attention was paid to this important sub-sample of the reviewed studies). We appreciate that, with a set of studies as heterogeneous as those included in this review, there is likely no ideal way to present results, but we propose retaining our original structure to ensure that the breadth of research is effectively communicated, and that U.S.-based studies on cannabis regulation do not overwhelm the other important studies included in this review. With respect to the studies by Model and Pollini, we thank the Reviewer for highlighting this. We have added a mention that both Model and Shi found beneficial impacts on non-cannabis hospitalizations:	pg. 12, para 2

	Further, two studies that also examined changes in acute care use for non-cannabis drugs found reductions in those visits or admissions following cannabis decriminalization or legal regulation [60, 79]. We appreciate the Reviewer's highlighting of the research by Pollini et al., which finds different results across two different outcomes and data sources (roadside testing for THC versus testing of fatal crash victims). However, because our analysis is conducted at the outcome (rather than paper) level, and we have not otherwise described within-paper discrepancies in detail, we are hesitant to include more granular detail on this particular study. This is, again, related to ensuring that we are able to communicate a heterogeneous literature within a reasonable word count; if the Editor wishes, we would be happy to expand the discussion to include granular study details.	

The article still makes a rather arbitrary distinction between studies that are included or excluded on the basis of whether they are 'original'. It state, 'We did not consider as original research studies that reproduced secondary data without conducting original statistical analyses of the data.' As many of the included studies used secondary data, this leaves the application of statistical analysis as the definition of whether research counted as 'original'. But such statistical analysis is not always appropriate as a method for comparative, longitudinal studies of the effects of drug policy changes, as it requires a relatively large sample of cases and comparators. For example, the study by Hughes et al. of 'Inconsistencies in the assumptions linking punitive sanctions and use of cannabis and new psychoactive substances in Europe' includes a sample of only 7 countries. It would not have been appropriate to apply inferential statistical tests to this sample, yet the study provide valuable information on the effects of legal regulations on levels of drug use (or, rather, the lack of such an effect). Given that the USA is the only country in the world which systematically publishes data on drug prevalence across a reasonably large sample of geographic units, it is unsurprising that this definition of 'original' research has excluded much research from other countries. There are studies that use statistical analysis of data from large cross-national surveys (e.g. ESPAD, HBSC) on the effects of national policy regimes on drug use prevalence, but these have not	We thank the Reviewer for this important comment. In designing inclusion and exclusion criteria for our review we applied the epidemiologic hierarchy of evidence; we appreciate that this standard privileges research from settings where data on drug use and related harms are longitudinally collected and made available. We note that highlighting the limitations of current methodological approaches to questions of regulation and decriminalization is a key motivator for this study. In response, we have substantially revised the discussion of the limitations section to more explicitly address how our methodologic inclusion criteria favor US studies; to provide examples of excluded studies; and to recommend future reviews inclusive of a wider range of evidence types. Further, while we are not currently resourced to undertake a novel review that employs different inclusion criteria, we do hope to continue work in this area and to motivate others to do so:	Pg. 15, para 2; pg. 14, para 2

been included in this review due to their cross-sectional design.	In addition, we excluded grey literature, non-original research, and study designs that are not suited to evaluating policy effects (e.g., cross-sectional studies), but these restrictions ~~may have~~ narrowed the geographic scope of included studies. **For example, two articles on Portugal were excluded as non-original research, but nevertheless provide important insight on impacts of decriminalization [99,100]. Despite restricting eligibility to more rigorous study designs,** most included studies used **relatively** weaker eligible designs that are **known to be** vulnerable to pre-existing trends and confounding; only 22.8% and 5.3% respectively used controlled before-and-after or interrupted time series designs to address these threats to validity. The use of these study designs may be related to limited resources for prospective drug policy evaluations, with many studies relying on publicly available, routinely collected data. **That the U.S. is unique in the extent to which data on drug use and related harms are routinely collected helps to explain its over-representation in our review. Scoping reviews inclusive of grey literature and cross-sectional designs would be valuable for describing the full range of evaluations that have been conducted globally.** We also clarify that “peer-reviewed **longitudinal** evaluations of drug decriminalization and legal regulation are concentrated in the U.S.”	

The article still does not acknowledge the contribution or existence of some previous reviews on the topic. It uses the reviews by Dirisu, Sarvet, Melchior and their co-authors to make the point about the narrow focus of previous reviews on adolescent cannabis use. However, some of the other reviews (as cited in my previous review of this article) have a broader focus than this.	As is common in setting the context for a systematic review, we were specifically assessing the scope of peer-reviewed systematic reviews. We did review each of the recommended citations, but the others were non-systematic literature reviews (narrative reviews) or other types of research. We have however revised the sentence citing those reviews to read: **Systematic reviews,** including two meta-analyses, are narrowly focused on adolescent cannabis use.	Pg. 4, para 3
France has (correctly) been removed from the list of countries that have adopted decriminalization. However, Belgium is still included, despite still using criminal penalties for possession.	We defined decriminalization as the "removal of criminal penalties for drug use and/or possession (*allowing for civil or administrative sanctions*)". We include fines under the definition of civil or administrative sanctions. As initially noted by the reviewer, Belgium replaced prison sentences with fines. We note that this distinction aligns with classifications employed by EMCDDA.	

VERSION 3 – REVIEW

REVIEWER	Alex Stevens University of Kent, UK
REVIEW RETURNED	01-Jul-2020

GENERAL COMMENTS	Sufficient changes have been made to warrant publication. The article will be a useful addition to the literature.