The Editorial Board Member for this manuscript was Joe Ross.

Comments from the Editorial Board Member and Reviewers:

Editorial Board Member:

A well-written article describing a database that will be highly useful for critical care medicine research. We are interested in a revision of your article, responding to the excellent reviewers comments that suggest a number of ways the manuscript should be clarified and strengthened prior to publication.

# 

# Reviewer #1

The paper "The eICU Collaborative Research Database, a freely available multi-center database for critical care research" describes the eICU Collaborative Research Database (eICU-CRD), which is a database with high granularity multi-center data for over

200000 admissions to Intensive Care Units (ICUs) in the United States monitored by the eICU Programs by Philips Healthcare.

In the 'Background & Summary' section, you refer to the critical care history and the relevance of data in this domain. You describe the Philips Healthcare eICU program and how the eICU Research Institute (eRI) transforms the data archived by Philips into a database, and how the MIT Laboratory for Computational Physiology (LCP) partnered with the eRI to produce the eICU-CRD.

The database website (http://eicu-crd.mit.edu) is not referred anywhere in the manuscript and I suggest that you add it, either in the introductory section, in the 'Data access' section or both.

We referenced the documentation under the “Collaborative Code and Documentation” section as follows: “Detailed documentation is also available online \cite{eicu-website} and is continuously updated.”. However, the editor has highlighted to us that the appropriate method for dynamic websites linked to by URLs is to embed them in text itself (which is also much clearer to the reader as you point out). The current reference, [3] in the draft you reviewed, has been updated from physionet.org and replaced with a footnote which redirects to <https://eicu-crd.mit.edu>. In general we have cleaned up our web presence thoroughly (which we discuss in more detail in reply to your comments on the Code Availability section).

LCP has previously shared the MIMIC-II and MIMIC-III databases, which as opposed to the eICU-CRD database that is multi-center, contain critical care data from a single center. You indicate that eICU-CRD builds upon the success of MIMIC-III and addresses its limitations. I would like to see a clarification on how the eICU-CRD database relates to the MIMIC-III database. Is there any overlapping data between the two resources, and the MIMIC-III single center was considered as part of the group of centers included in eICU? If not, why not?

eICU-CRD only contains hospitals which license the Philips eICU programme. The source hospital of MIMIC-III, the BIDMC, does not license this programme and therefore is not in eICU-CRD. We have added the following sentence in the introduction to clarify this: “As the source hospital of MIMIC-III does not participate in the eICU programme, eICU-CRD is a completely independent set of data collected from a large number of hospitals located within the United States.”

What are the limitations you identified for MIMIC-III (in addition to being data about a single-center) and how are these limitations addressed? The manuscript does not describe these aspects at the moment.

The primary limitation, and the one we would like to focus on in this manuscript, is the sourcing of MIMIC-III from a single center. This is consistently raised as an issue in manuscripts generated from MIMIC-III, and is the biggest advantage of eICU over MIMIC-III. We have rewritten the end sentence of the introduction to focus on this: “The release of eICU-CRD is intended to build upon the success of MIMIC-III and expand the scope of studies possible by making data available from multiple centers.”

In the 'Methods' section, under 'Database structure and development', you describe the eICU-CRD relational database, as comprising of 17 tables. Is the relational model available for potential users of the data? Where? The relational model/schema is important for users to be able to query the database, so I suggest it is made available.

We have made the relational model available online using an open source tool called SchemaSpy. The schema is available here: <https://mit-lcp.github.io/eicu-schema-spy/>

This schema is linked to via the main documentation website.

It’s worth noting that eICU-CRD is very much denormalized, and so the almost every relation is between the patient table and other tables. We have highlighted this in the manuscript.

You refer to the database's version: v1.2 - is this version referring to the data that you are making available or to the database model? Other references to this version number are linked to the list of tables in the database, so if referring to the database model, do you foresee future changes to the model?

The aim of the versioning system was to use major version changes for schematic changes which syntactically break code, and use minor version changes for updates to the data (i.e. we hope to adhere to semantic versioning principles). We do not foresee many changes to the data model, and are primarily versioning the data available. It is possible that issues are raised in the future by users of the data which we must correct, Similar to the MIMIC changelog (<https://mimic.physionet.org/about/releasenotes/>), we will highlight any changes we make to the data for each release. Finally, as the data extraction process done by Philips randomizes identifiers, future updates to the dataset (with more data e.g. from 2016-2018) would be given a new version number. We describe this in the manuscript in the “Data Access” section.

In undertaking this review, we have made a substantial update to the data which we discuss in our cover letter. Thus, the current version of eICU-CRD is now v2.0.

Table 4 (referenced in the section 'Classes of data') provides a list of the tables available in the eICU-CRD (v1.2), so it would be useful to refer to Table 4 in this section ('Database structure and development') describing it. In fact, the text refers to 17 tables, but Table 4 lists 20 tables - why is there a discrepancy here?

In addition, the sample data that you provided includes 21 datasets, most of which correspond to the 20 tables listed in Table 4, except for allergy - is this an additional table that was not included in the listing?

On further investigation, by looking at the Jupyter notebooks on GitHub, I came across the documentation about the tables at http://eicu-crd.mit.edu/eicutables/admissiondrug/

Again, I find more tables listed that are not in Table 4.

Apologies for the discrepancy. We are continually updating eICU-CRD, and missed inconsistencies in the manuscript. Online, we attempted to document all tables which we foresee releasing in the future, but in the paper we only described currently released tables. In performing these revisions, we acknowledge that it was an oversight to not also include these tables in the manuscript. In an internal discussion, we agreed that it would be best if the paper described all tables planned for eICU-CRD, and that the best mechanism for accomplishing this would be to release data for all tables. Thus, with some effort, we have released a new version of eICU-CRD with data from all tables. We have thoroughly updated the manuscript for consistency and the tables listed in the data are the final set of tables planned for eICU-CRD.

In addition, the access to this documentation should be more straightforward for any reader of the manuscript and any user of the data. Please, make it more visible. As mentioned above, I also suggest that you make available the relational model, otherwise the list of tables is insufficient to understand how the tables (and thus the data) are related.

We have added the documentation as a footnote in the Collaborative code and documentation section, which hopefully increases the visibility. We include the relational model in this documentation and highlight this in the section with the following sentence: “The documentation includes a schematic of the relationships between tables in the data.”

It is indicated that the data was de-identified to comply with the US HIPAA and that large portions of the tables were reviewed manually by at least three personnel. What method was used for de-identification, what data elements were removed, how did you deal with dates, etc?

Please, provide more details on the de-identification process. Have you measured the probability of re-identification (see e.g. <http://doi.org/10.1136/bmj.h1139>)?

As much as possible, we have added detail on the deidentification process. The deidentification proceeded in three stages: (1) a commercial company, PrivaCert, was contracted and annually reviews the data schema to ensure it sufficiently meets both statistical and safe harbor standards established by HIPAA, (2) the data extraction process by Philips medical removes all sources of expected PHI as proposed by PrivaCert (e.g. free-text fields), and finally (3) fields which are not sources of expected PHI are further analyzed and PHI is removed as needed. We have added the following section to detail this process:

“A security schema was established, for which the re-identification risk was certified as meeting safe harbor standards by Privacert (Cambridge, MA) (HIPAA Certification no. 1031219-2).

Subsequent to this certification, free-text fields were scanned for PHI using a previously published rule-based approach (<https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/1472-6947-8-32>). Briefly, this approach scans text for known patterns indicating presence of PHI (e.g. words following ``Mr.'' are frequently names, such as ``Mr. Smith''). The approach also detects words which are commonly used as places or names. The output of this algorithm was reviewed, and rows containing PHI were deleted. Finally, large portions of all tables were manually reviewed by at least three personnel to verify all data had been de-identified. Frequently, due to a low number of unique entries (e.g. when a table stored the results of a drop down menu), the entire table was reviewed.“

It is indicated that the patient identifiers are generated randomly, and thus, the eICU-CRD information cannot be linked back to the original identifiable data. In the section on 'Patient Identifiers', you go into more detail to describe the different identifiers associated with a patient, their hospital admissions, and their ICU admissions. Figure 1 shows these different identifiers (but has an error on the patient level identifier that should be uniquePid instead of patientUnitStayId).

Apologies - we have corrected this.

The figure and the legend use camel case words, while the rest of the text doesn't. Please, change this to be consistent (while the camel case is more readable, I suggest to follow the way it is represented in the tables that is not camel case, you may find a way to make it more readable by using bold font instead of camel case for the first letter of each word, for example).

We have added a brief outline of our choice at the beginning of the methods section, as follows:

“Each file contains data for a single table, and we denote references to tables by using monospace font. Similarly, we denote references to columns using *italicized font*. To enhance readability we use camel case where possible, though all files and columns are distributed using lower case letters.”

We hope this format facilitates readability.

A section on how a stratified random sample of patients was selected is included - what was the use of this sample? Is it also available for re-use, where/how?

eICU-CRD is a subset of data archived and utilized by Philips Healthcare. This section describes the sampling approach. We have introduced this section differently to make it clearer, as follows: “eICU-CRD is a subset of a research data repository maintained by eRI.

A stratified random sample of patients was used to select patients for

inclusion in eICU-CRD...”.

The section 'Code availability' lists two GitHub repositories:

- https://github.com/MIT-LCP/eicu-website - the website and documentation for the eICU-CRD

- https://github.com/MIT-LCP/eicu-data-paper - which actually doesn't exist.

It would be useful to list the main website where people should be able to get information about the eICU-CRD as a single point of access: http://eicu-crd.mit.edu

In addition, the inexistent GitHub repository may now be: https://github.com/MIT-LCP/eicu-code that should be included.

A reference to the material listed at https://github.com/MIT-LCP/ACM-BCB-2017 should be useful, as it is indicated that corresponds to a tutorial at ACM-BCB-2017 about the eICU-CRD.

Indeed, our web presence was fragmented. We have made the following changes:

1. eicu-data-paper - this repository existed, but was private. We have now made it open.
2. eicu-website - this repository has been merged into eicu-code, for simplicity
3. eicu-code - this is now referenced in the paper
4. ACM-BCB-2017 - this is likely to become out of date (and may already be out of date with v2.0), so we would prefer not to mention it in the paper, but link to it via eicu-code. We plan to keep eicu-code up to date with tutorials around the data.

The online documentation now also embeds Jupyter notebooks from eicu-code which overview usage of each table in eICU-CRD.

In the section on 'Data Records', you provide a high level summary of the database information. Table 1 provides summary level demographic information - is the code generating this table available on the GitHub repository? If so, please include a reference associating the notebook to the Table. If not, why not?

Similarly, it would be good if you could provide the code generating the other summary tables (Table 2 on most frequent admission diagnoses, Table 3 on most frequent categories of APACHE diagnosis using clinically meaningful groups, Table 5 for the percentage of hospital information according to different categories, etc.).

The code to generate these tables is available in the eicu-data-paper repository. We have added a reference to this repository using Zenodo to generate a DOI. This reference is provided in the Code Availability section.

In the sub-section on 'Care documentation', you indicate that treatments are documented using a hierarchical coding system. What is this system? Is this a standard one?

This is a custom hierarchy defined by Philips. There is no standardized ontology for these treatments, though a fraction could foreseeably be mapped to a standard ontology, which we hope the community endeavors on.

Users may find it useful if you provide a list of the variables kept in the database. This could be in the form of online documentation that you could link in the data descriptor.

This is very difficult to do due to the large number of variables available. Furthermore, as the exact variables available for each hospital will vary, we believe it would be misleading to create a page which lists “variables available”. However, as much as possible, the variables are described in the table specific usage notebooks now provided on each table page of the online documentation.

When describing the 'Technical validation', you indicate that "data were verified for integrity during the data transfer process from Philips to MIT" - do you mean by checksums or what other verification method?

We now explicitly state that we used MD5 checksums.

You mention there is public issue tracker used as a forum for reporting technical problems and discussing solutions - is this the GitHub issue tracker? Of which of the two repositories? Please, include the link and suggestion of usage.

We plan to use the issue tracker in eicu-code as the primary forum for discussing use of the data. In the Usage section (subsection Collaborative code and documentation), we now explicitly state the use of the issues: “Discussion around usage of the data, highlighting of issues, and inquiry regarding best practices can be made via the issues panel of the GitHub repository\footnote{https://github.com/MIT-LCP/eicu-code/issues}.”

Related to the "Usage Notes", the text points to the PhysioNetWorks repository for access to the data but at the moment the eICU-CRD is not linked from the main PhysioNet page: https://physionet.org/ - why not refer to the main eICU-CRD website http://eicu-crd.mit.edu ? Actually the website in reference [3] (https://eicu-crd.physionet.org/) is down.

We now highlight that data access steps are described on the eicu-crd.mit.edu website, and provide it as a footnote.

Regarding the data use agreement, I would expect to see one specific for eICU-CRD, as currently you provide the text available for the PhysioNet Clinical Database.

We only have one DUA for all restricted clinical databases which we provide. We do not require any provisions unique to any particular dataset, and so we plan to use a single DUA. Those who are approved for access can utilize MIMIC-III, or eICU, or both.

In the 'Example usage' section, it would be useful to provide a specific link to the Jupyter notebook. In fact, I recommend you make a release of the repository and associate it with a DOI in Zenodo, so that you can provide an appropriate software citation, ensuring that future changes in the GitHub repository won't affect this citation

(see https://guides.github.com/activities/citable-code/).

This is actually exactly what we intended to do! We have created DOIs with Zenodo for both eicu-code and eicu-data-paper. However, as far as we are aware, we cannot link to particular aspects of an object with a DOI, only to the entire object. Thus, we have cited the repository and rephrased the sentence to make it obvious we refer to the only Jupyter Notebook available in that repository: “Figure \ref{fig:example\_patient} provides a visualization for a subset of

the patient's stay, and the Jupyter notebook used to generate this figure is available

online \cite{eicu-data-paper}.”

The quality of Figure 2 should be improved, adding labels to the axis, etc.

We have remade Figure 2 with a y-axis label and changed the patient to one present in the demo database. We have also changed the plotted data elements to reflect new tables available.

Some issues I see with the data:

- Semantics and standardization

Currently, the data lacks semantic annotations, which are required for enabling data interoperability, integration and data re-use at scale.

While APACHE IV is not linked to a terminology, there have been approaches providing mappings (e.g. for a mapping to SNOMED-CT see https://www.ncbi.nlm.nih.gov/pubmed/18487826).

Given that the MIMIC-III database relies on the International Classification of Diseases (ICD-9), it would be useful that the same coding system is used in the eICU-CRD. Do you have suggestions, or even better, a computational method to enact such integration between the two resources?

Thank you for the link to the paper mapping APACHE IV diagnoses to SNOMED-CT. We very much agree that a consistent ontology would greatly enhance the database. Unfortunately, this is an incredibly labor intensive task. Our hope is that this will become a community effort, and that as projects are completed, the relevant variables are standardized and the code for this standardization is made available to the community. Perhaps, during this effort, computational tools will be developed which can help. For example, the USAGI tool (<https://github.com/OHDSI/Usagi>) is one such tool which aims to facilitate concept mapping. However, it uses simple fuzzy string matching, and so strings such as “ATach” and “VTach” are considered very similar, while “Ventricular Tachycardia” and “VTach” are considered very different. Again, the hope is that these tools will be developed by the community over time, perhaps using eICU-CRD as an exemplar.

On a related note, there are ICD codes available for the diagnosis table, and we have added this detail to the manuscript.

Similarly, the database has a custom schema and thus, custom data processing and analysis methods need to be develop. The Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) is being increasingly used and following this common model will facilitate integration with MIMIC-III (given its existing transformation to OMOP-CDM https://github.com/MIT-LCP/mimic-omop). Could you provide a similar transformation from eICU-CRD?

This is certainly a research direction we think is very interesting. While the OMOP CDM is becoming increasingly used, it can be challenging to integrate EHR data into the model, which we found when mapping MIMIC to OMOP. Similar to our approach for MIMIC-III, we hope to publish the data as is, and focus future research efforts on moving the data into a common data model.

- Provenance information about the data:

Across the data descriptor, there is information about the provenance of the data: e.g. you indicate that the hospital information is the result of a survey, and there is some implication that this is the reason for the data to be incomplete. Also, the data in the tables 'vitalperiodic' and 'vitalaperiodic' contains data derived directly from the bedside monitors, with no validation by a clinician. On the other hand, other data within the database has human verification, which ones?

All other tables have had human verification. We have clarified this as follows: “Unlike other data elements in the database, the data collected in these tables are not entered or validated by providers of care; i.e. the periodic and aperiodic vital sign data have been automatically derived and archived with no human verification.“

To facilitate data re-use with some confidence on the conclusions that could be derived from the data, it would be important to record as part of the database itself this provenance information, i.e. data about the origin of the data, indicating that the hospital data comes from a survey, and what information is or not verified by a clinician. There are models allowing such a representation (e.g. see the provenance ontology).

Where possible, we have described the provenance of the data within the paper. While explicitly documenting this using an ontology and a new data model would be useful for many applications, we do not have the ability to provide this information.

- Units of measurements

The data tables do not include information about the units of measure (e.g. in the 'patient' table, there is no information about the unit for 'admissionweight'). While this information is available in the documentation page http://eicu-crd.mit.edu/eicutables/patient/ -- that not listed in the manuscript -- it should also be added to the dataset, considering their semantic representation to enable interoperability (e.g. see <https://www.ebi.ac.uk/ols/ontologies/uo>)

We agree it is a limitation of the data model to not store unit of measurement appropriately. However, we aim to maintain the data model as much as possible as it is desirable for Philips Healthcare and as we will likely pursue mapping the data to OMOP CDM in the future. We do not believe improving the data model is worth the effort; rather this effort would be better spent converting the data to something like the OMOP CDM.

The instructions given to access the data using the sqlite file are dependent on the version of Firefox used, as the SQLite Manager add-on does not longer work in the latest Firefox version (Firefox Quantum). Thus, the instructions need to be updated to indicate the Firefox version to use (that is not the latest), and/or other ways of installing sqlite should be given for different platforms.

Unfortunately we were only made aware of this after we submitted the data instructions, apologies for any debugging you had to do.

It would be important to include details on subsequent versions of the database: are you planning to add more data and provide other releases in the future? Please, provide details on the sustainability of the resource.

We do plan on providing updates, and have added the following text to the data access section: “Finally, eICU-CRD v2.0 contains data for patients admitted between 2014-2015. Future updates will be made to ensure the data remains contemporary.”

Other minor issues Please proof-read the article again and correct the syntactic issues, for example:

- in the 'Background & Summary' section: "...with deteriorating illness athat trigger reassessment of the treatment regimen..." - remove 'a'

- in the section on 'Patient identifiers', "Each patient stay in a unit...", fix 'stay'

Corrected, thanks.

Please, enumerate the tables in their order of appearance in the manuscript.

We have reviewed the tables and ensured they are labeled correctly. The order of Table 4 is now alphabetical for clarity.

Overall, I find that the eICU-CRD has a lot of potential to be a great resource but:

- the data descriptor needs to provide more details on the database (e.g. harmonize details about the tables, provide schema, more details on the de-identification process, and so on as above), as well as the links to access the database and documentation need to be harmonized and fixed, figures improved, etc.

We hope that we have satisfactorily addressed all these concerns.

- the database itself could be improved in terms of defining the provenance of the information, as it varies greatly across tables, the unit of measurement should be added to the database with semantic annotations, semantic annotation of other aspects (e.g. diagnosis) should be added, a mapping to common models such as OMOP-CDM should be considered. Semantic and standardization considerations would enable integration with other data such as MIMIC-III, making the data re-use at scale more feasible.

While we strongly agree that these are key steps towards improving the usability of the data, we would argue that they are out of scope. The OMOP CDM is a laudable initiative, but the translation of a high resolution database such as eICU-CRD into OMOP CDM is still an area of active research. Our laboratory plans to pursue this field of research to the fullest in the years to come.

# Reviewer #2

To the editor,

Thank you for your invitation to review the manuscript "The eICU Collaborative Research Database, a freely available multi-center database for critical care research". Alistair Johnson et al. described a multi-center database with high resolusion data during ICU stay for more than 200,000 patients monitored by eICU Programs across the United States. This database included very detailed information supporting further clinical research in these patients. More importantly, this database is freely open to researchers who passed the specific training course, which could promote the development of big data research in critical care medicine.

However, several limitation need to be addressed.

Experimental Rigor and Technical Data Quality

All these data were extracted directly from the eICU program electronially, and was presented as raw data in this database, with very little revision during post-processing. Thus, the information bias during the processing was largely minimized and the reliability of these data was relatively high. However, the limitation need to be noticed especially in tables without clinician validation. For instance, the data recorded in vitalperiodic table are recorded every 5 minutes automatically, without validation. Thus, incorrect values such as blood pressure are also recorded and may be difficult to distinguish.

We have added a sentence to highlight that all data, aside from the vitalperiodic and vitalaperiodic table, have been manually recorded by healthcare professionals.

Completeness of the Description

This publication of this database aims to support a number of investigation in critical care area. Personally, I’m little confused by the abbreviations of items in tables. For example, the meaning of “interventioncategory” in table carePlanCareProvider, the “cplgroup” in table carePlanGeneral. A detailed explaination of items maybe helpful for clinicians.

We tried to strike a balance in the paper between high level information and low level column description. Our choice was to document the details of individual columns in the online documentation provided at <http://eicu-crd.mit.edu> - on this website, each table has its own page and a description of each column is provided. We hope that this will facilitate use of the data, and we further hope that users will raise issues at the GitHub repository hosting the data (<https://github.com/MIT-LCP/eicu-code/issues>) to highlight deficits in the documentation and collaboratively improve it over time.

On the other hand, I noticed there is a “diagnosisid” in table diagnosis, is there a de-code table for the “diagnosisid”?

In fact, this column has no meaning, and is merely used to prevent duplicate row insertion in a database. We have added this information to the manuscript: “All tables, other than patient and hospital, have a randomly generated primary key with the suffix `id', e.g. the diagnosis table has diagnosisId as a primary key. This column has no physical meaning, and is used to constrain uniqueness on rows and ensure integrity of the data when loading into a database system.”

For lab items, 158 distinct types of laboratory measurements are available. However, in the demo, 5% patients have less than 32 lab records. For instance, there are only 9 records for patients 222418 during the whole ICU stay. Is this because this is a demo or the true number of records of these patients is relatively small?

The number of lab records available for each patient is variable, and depends on their hospital course. What we refer to by “158 distinct types” is the different variables available in the table which can be broadly classified as “labs”: creatinine, troponin, sodium, and so on. We have clarified this in the text.

Integrity of the Data Files and Repository Record

A demo including 1,000 patients randomly selected from the full dataset was supplied for review. The selection process is strict and is fully described in the manuscript. The data files in this demo are complete and match the descriptions in the Data Descriptor. One issue should be explained. Mortality is an important outcome in CCM. According the description of the demo, all the patients in the demo were eventually dead. However, we noticed in Table “patients”, 283 of 1520 patients were alive according the hospitaldischargestatus. Is there other record for mortality?

These patients eventually died but had multiple hospitalizations. Therefore, 283 of the 1520 ICU stays had a hospital discharge status of alive. Grouping by *uniquePid*, and selecting the last discharge status, would reveal that all patients in the demo died.

To simplify matters, we have recreated the demo with a random sample of patients, both alive and dead. This also makes the demo much more useful for demonstrations.

The following two websites in the manuscript are not active. <https://github.com/MIT-LCP/eicu-data-paper/>

<http://eicu-crd.physionet.org>

We have reorganized our web presence, and now have the following repositories/websites:

1. eicu-data-paper - this repository existed, but was private. We have now made it open.
2. eicu-website - this repository has been merged into eicu-code, for simplicity
3. eicu-code - this is now referenced in the paper

We have added footnotes and/or references to these repositories as appropriate (Scientific Data encourage footnotes for dynamic web pages and references for static web pages).

Overall, this is a great database for critical care medicine. Of course, efforts are still needed for further improvement and update.

Many thanks - and we do hope to continually iterate upon and improve the data both during and after publication.