

OPEN  
ARTICLE

# Rethinking clinical study data: why we should respect analysis results as data

Joana M. Barros<sup>1,2,3</sup>✉, Lukas A. Widmer<sup>1</sup>, Mark Baillie<sup>1</sup>✉ & Simon Wandel<sup>1</sup>

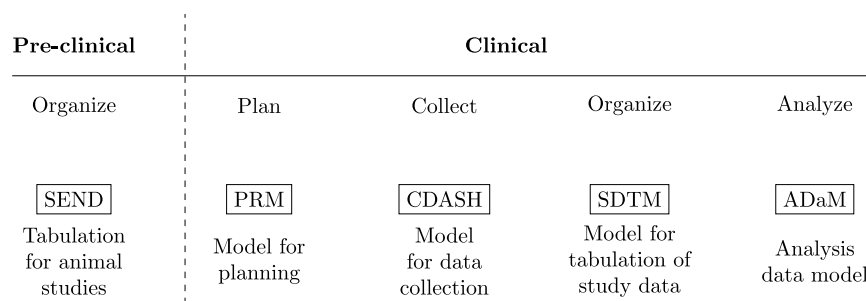
The development and approval of new treatments generates large volumes of results, such as summaries of efficacy and safety. However, it is commonly overlooked that analyzing clinical study data also produces data in the form of results. For example, descriptive statistics and model predictions are data. Although integrating and putting findings into context is a cornerstone of scientific work, analysis results are often neglected as a data source. Results end up stored as “data products” such as PDF documents that are not machine readable or amenable to future analyses. We propose a solution to “calculate once, use many times” by combining analysis results standards with a common data model. This analysis results data model re-frames the target of analyses from static representations of the results (e.g., tables and figures) to a data model with applications in various contexts, including knowledge discovery. Further, we provide a working proof of concept detailing how to approach standardization and construct a schema to store and query analysis results.

## Introduction

The process of analyzing data also produces data in the form of results. In other words, project outcomes themselves are a data source for future research: aggregated summaries, descriptive statistics, model estimates, predictions, and evaluation measurements may be reused for secondary purposes. For example, the development and approval of new treatments generates large volumes of results, such as summaries of efficacy and safety from supporting clinical trials through the development phases. Integrating these findings forms the evidence base for efficacy and safety review for new treatments under consideration.

Although integrating and putting scientific findings into context is a cornerstone of scientific work, project results are often neglected or indeed not handled as data (i.e., the machine-readable numerical outcome from an analysis). Analysis results are typically shared as part of presentations, reports, or publications addressing a greater objective. The results of data analysis end up stored as *data products*, namely, presentation-suitable formats such as PDF, PowerPoint, or HTML documents populated with text, tables, and figures showcasing the results of a single analysis or an assembly of analyses. Contrary to data which can be stored in data frames or databases, data products are not designed to be machine-readable or amenable to future data analyses. An example comparing a data product with data is given in Fig. 3. In this example, we illustrate how a descriptive analysis of individual patient data - in this case the survival probability by treatment over time - then becomes a new machine-readable data source for subsequent analyses. In other words, the results from one analysis becomes a data source for new analyses. This is the case for clinical trial reporting where the data analysis summaries from a study are rendered to rich text format (RTF) files that are then compiled into appendices following the International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH) E3 guideline<sup>1</sup> where each appendix is a table, listing or and figure summary of a drug efficacy and safety evaluation. The analysis results stored in these appendices - which can span 1000s of pages - are not readily reusable: extracting information from PDF files is notoriously difficult, and even if machine-readable formats (RTFs) are available, often some manual work is required since important (meta-)information is contained in footnotes for which no standard formats exist. There have been recent attempts to modernise the reporting of clinical trials including the use of electronic notebooks and web-based frameworks. However, while literate programming documents such as Rmarkdown allow documenting code and results together and R-shiny enables dynamic

<sup>1</sup>Analytics, Novartis Pharma AG, Basel, Switzerland. <sup>2</sup>Department of Biometry, Idorsia Pharmaceuticals, Allschwil, Switzerland. <sup>3</sup>This work took place and was submitted when the author was at Novartis: Joana M. Barros. ✉e-mail: joanamarquesbarros@gmail.com; mark.baillie@novartis.com



**Fig. 1** CDISC defines a collection of standards adapted to the different stages in the clinical research process. For example, ADaM defines data sets that support efficient generation, replication, and review of analyses<sup>36</sup>.

data exploration, the rendered data products also suffer the same fate of presentation-suitable formats. In other words, modern data products also do not handle data analysis results as data. Although there is an agreement on which information should be shared as part of a data package and that sharing data can accelerate new discoveries, there is no proposed solution to facilitate the sharing and reuse of analysis results<sup>2</sup>.

A focus on results presentation over storage considerations sets up a barrier impeding the assimilation of scientific knowledge, understanding what was intended and what was implemented. As a repercussion, the scientific process cycle is broken, leaving researchers who want to reuse prior results with three options:

1. Re-run the analysis if the code and original source data are accessible.
2. Re-do the analysis if only the original source data is accessible.
3. Manually or (pseudo-)automatically extract information from the data products (e.g., tables, figures, published notebooks).

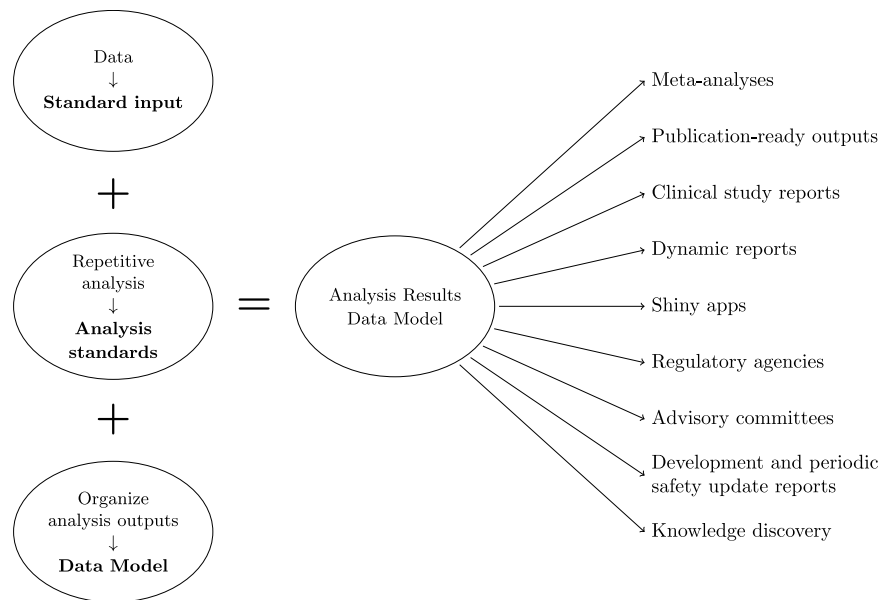
The first option would appear to be the best one and is, for instance, being implemented in Elife executable research articles<sup>3</sup>. However, being able to rerun the analysis does not guarantee reproducibility and can be computationally expensive when covering many studies, large data, or sophisticated models. Analyses can depend on technical factors such as the products used, their versions, and (hardware and software) dependencies, all of which affect the outcome. Even tailored statistical environments such as R<sup>4</sup> have a wide range of output discrepancies and must rely on extensions, such as *broom*<sup>5</sup> for reformatting and standardizing the outputs of data analysis.

For the second option, there are additional complications to account for: even if we assume that the entire analysis is fully documented, common analyses are not straightforward to implement. This option assumes that the complete details required to implement the analysis are documented, for example, in a statistical analysis plan (SAP). However, data-driven and expertise-driven undocumented choices are a hidden source of deviations that make reproducing or replicating the results an elusive task<sup>6</sup>. On top of this, the selective reporting of results limits replication of the complete set of performed data analyses (both pre-specified and ad-hoc) within a research project<sup>7-9</sup>.

The last scenario is common place for secondary research that combines and integrates findings of single, independent studies, such as meta-analyses or systematic reviews. Following the Cochrane Handbook for Systematic Reviews of Interventions to perform a meta-analysis, to assess the findings, it is necessary to first digitize the studies' documents either through a laborious manual effort or by using extraction tools known to be error-prone and requiring verification<sup>10</sup>. Furthermore, the unavailability of complete results, potentially through selective reporting, requires researchers to extrapolate the missing results, which can lead to questionable reliability and risk of bias<sup>11</sup>.

Data management is an important, but often undervalued, pillar of scientific work. Good data management supports key activities from planning and execution to analysis and reporting. The importance of data stewardship is now also recognized as an additional pillar. Good data stewardship supports activities beyond the single project into areas such as knowledge discovery, as well as the reuse of data for secondary purposes, to other downstream tasks such as the contextualization, appraisal, and integration of knowledge. Initiatives like FAIR set up the minimal guiding principles and practices for data stewardship based on making the data Findable, Accessible, Interoperable, and Reusable<sup>12</sup>. Likewise, the software and data mining community (e.g., IBM, ONNX, and PFA) have introduced initiatives bringing standardization to analytic applications, thus facilitating data exchange and releasing the researcher from the burden of translating the output of statistical analysis into a suitable format for the data product.

An important component of data management is the data model which specifies the information to capture, how to store it, and standardizes how the elements relate to one another. In the clinical domain, data management is a critical element in preparing regulatory submissions and to obtain market approval. In 1999 the Clinical Data Interchange Standards Consortium (CDISC) introduced the operational data model (ODM) facilitating the collection, organization, and sharing of clinical research data and metadata<sup>13</sup>. In addition, the ODM enabled the creation of standards (Fig. 1) such as the Standard Data Tabulation Model (SDTM) and the analysis data model (ADaM) to easily derive analysis datasets for regulatory submissions. Owing to the needs at the different stages of the clinical research lifecycle, CDISC data standards reflect the key steps of the clinical data lifecycle. Although regulatory procedures were traditionally focused on document submission, there has since been a gradual desire to also assess the data used to create the documents<sup>14</sup>. CDISC data standards address this need; however, these standards only consider data from planning and collection, up to analysis data (i.e. data prepared and ready for data analysis). Therefore, the outcome of this paper can be viewed as a potential



**Fig. 2** In clinical development, the analysis results data model enables a source of truth for results applied in various applications. Currently, the examples on the right require running analyses independently, even when using the same results.

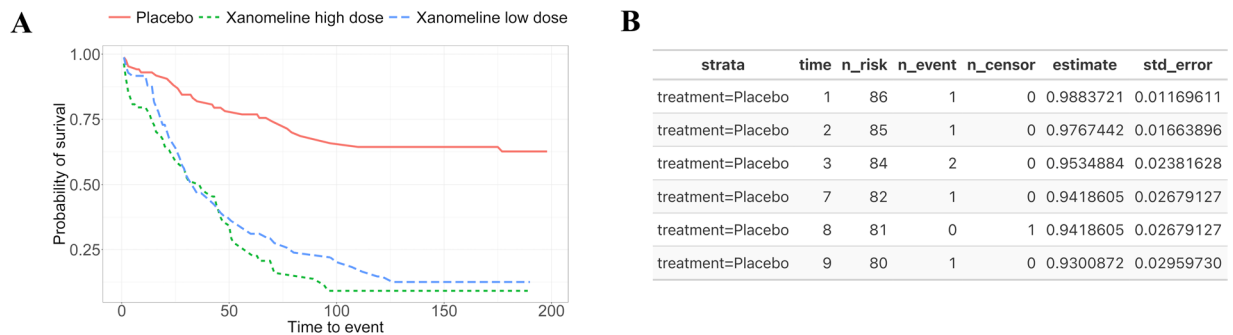
extension to the CDISC data standards and how not only individual patient data but also descriptive and inferential results should be stored and made available for future reuse.

In this paper, we explore the concept of viewing the output of data analysis as data. By doing so, we address the problems associated with the limited reproducibility and reusability of analysis results. We demonstrate why we should respect analysis results as data and put forward a solution using an *analysis result data model* (ARDM), re-framing the analyses target from the applications of the results (e.g., tables and figures) to a data model. By integrating the analysis results into a similar schema with specific constraints, we would ensure analysis data quality, improve reusability, and facilitate the development of tools leveraging the re-use of analysis results. Taking meta-analyses again as an example, applying an ARDM would now only require one database query instead of a long process of information extraction and verification. Tables, listings, and figures could be generated directly from the results instead of repeating the analysis. Furthermore, storing the results as independent datasets would also allow sharing information without the need for the underlying individual patient data, a useful property given data protection regulations in both academic and industry publications. Viewing analysis results as a data source moves us from repeating or redundantly recording results to a *calculate once, use many times* mindset. While we use the latter term focusing on results of statistical analyses for clinical studies, it can be seen as a special case of the more general concept of open science and open data, which aims at reducing redundancy in scientific research on a larger scale.

## Results

**Implementing the ARDM in clinical research.** The ARDM is adaptive and expandable. For example, with each analysis standard, we can adapt or create new tables to the schema. With respect to the inspection and visualization of the results, there is also the flexibility to create a variety of outputs, independent of the analysis standard. The proof of concept for the ARDM is implemented using the R programming language and a relational [SQLite database](#); however, these choices can be revisited as the ARDM can be implemented using a variety of languages and databases. This implementation should be viewed as a starting point rather than a complete solution. Here, we highlight the considerations we took to construct the ARDM utilizing three analysis standards (descriptive statistics, safety, and survival analysis) and leveraging the CDISC Pilot Project ADaM dataset. Further documentation is available in the code repository. An overview of the requirements to create the ARDM is shown in Fig. 2.

Prior to ingesting clinical data, the algorithm first creates empty tables with specifications on the column names and data types. These tables are grouped into metadata, intermediate data, and results. The metadata tables are created to record additional information such as variables types (e.g., categorical and continuous) and measurement units (e.g., age is given in years). As part of the metadata tables, the algorithm also creates an analysis standards table requiring information on the analysis standard name, function calls, and its parameters. The intermediate data tables aggregate information at the subject level and are useful to avoid repeated data transformations (e.g., repeated aggregations) thus, reducing potential errors and computational execution time during the analysis. The results tables specify the analysis results information that will be stored. Note that the creation of the metadata, intermediate data, and result tables require upfront planning to identify which information should be recorded. Although it is possible to create tables ad hoc, a fundamental part of the ARDM is to generalize and remove redundancies rather than creating a multitude of fit-for-purpose solutions. Hence, creating



**Fig. 3** The Kaplan-Meier plot corresponds to a data product from a survival analysis (A). On the contrary, the data from the analysis is stored in a machine-readable format (B) allowing for updates to the Kaplan-Meier plot and for use in downstream analyses.

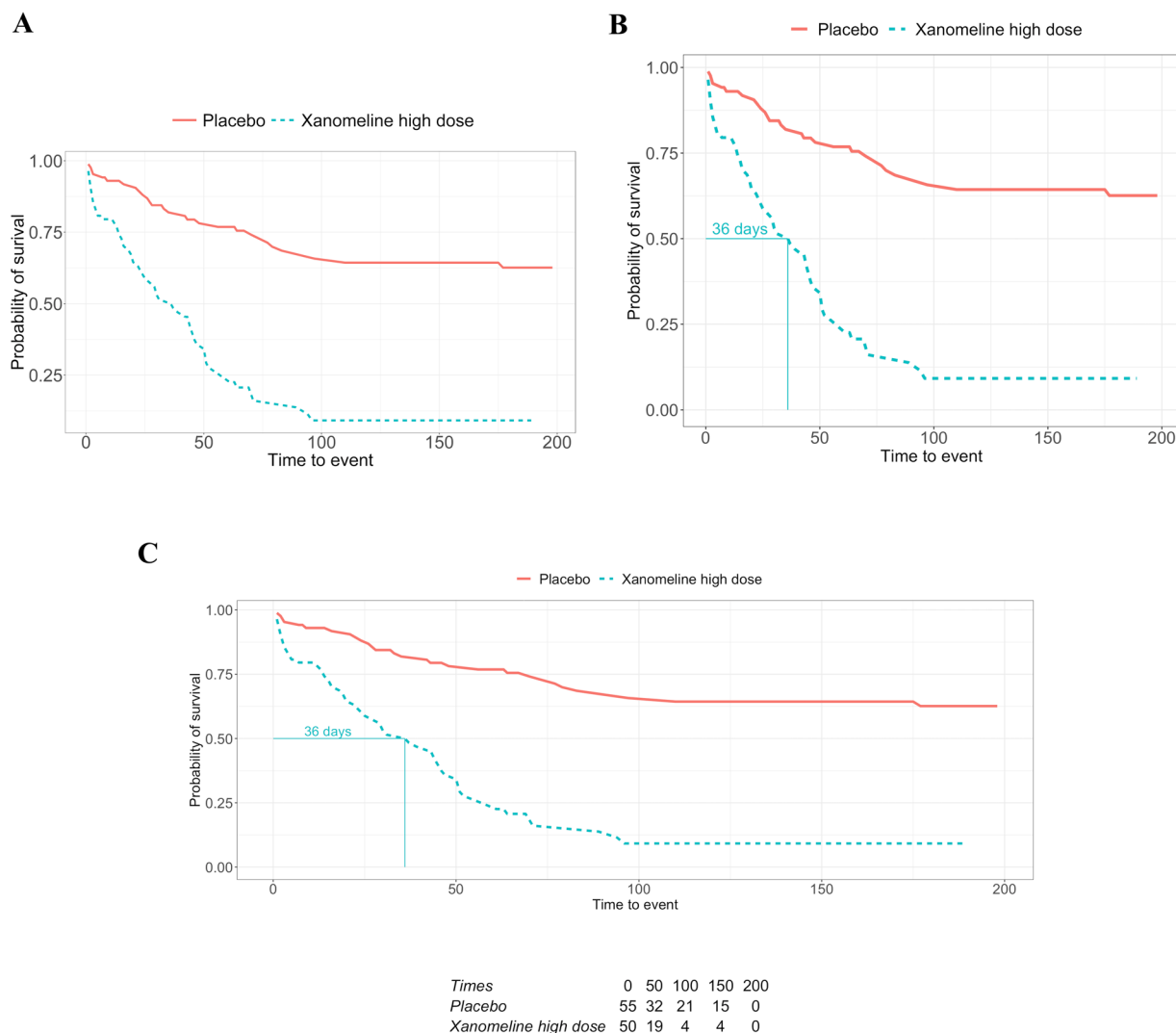
a successful ARDM requires understanding the clinical development pipeline to effectively plan the analysis by taking into account the downstream applications of the results (e.g., the analysis standard or the data products). As the information stored in the results tables is dictated by the data model, it is possible to inspect the results by querying the database and creating visualizations. In the public repository<sup>15</sup>, we showcase how to query the database and create different products from the results. Furthermore, the modular nature of the ARDM separates the results rendering from the downstream outputs hence, updates to the data products do not affect the results.

**Applications.** Analysis standards are a fundamental part of the ARDM to guarantee coherent and suitable outputs. They ensure that the results are comparable, which is not always the case. Similarly, where conventions exist (e.g., safety analysis), we can use an ARDM to provide structure to the results storage thus, facilitating access and reusability. In short, it provides a knowledge source of validated analysis results, i.e. a single source of truth. This enables the separation between the analysis and the data products, streamlining the creation of tables or figures for publications, or other products as outlined in Fig. 2.

Tracking, searching, and retrieving outputs is facilitated by having an ARDM as it enables query-based searches. For example, we can search based on primary endpoints “p-value”, “point estimates”, and adverse events incidence for any given trial present in the database. With automation, we can also select cohorts through query-based searches and apply the analysis standards to automate the creation of results using the selected data. This also facilitates decision-making and enhancements. For example, one can have access to complete trial results beyond the primary endpoint, and extrapolate to cohorts that require special considerations such as pediatric patients. In addition, a single source of truth for results encourages the adoption of more sophisticated approaches to gather new inferences, for example, using knowledge graphs and network analysis.

**Case study: updating a Kaplan-Meier plot.** The Kaplan-Meier plot is a common way to visualize the results from a survival or time-to-event analysis. The purpose of the Kaplan-Meier non-parametric method is to estimate the survival probability from observed survival times<sup>16</sup>. Note that some patients might not experience the event (e.g., death, relapse); hence, censoring is used to differentiate between the cases and to allow for valid inferences. As a result of the analysis, survival curves are created for the given strata. For the CDISC pilot study which was conducted in patients with mild to moderate Alzheimer’s disease, a time-to-event safety endpoint, the time to dermatologic events, is available. Such time-to-event safety endpoints are not uncommon in practice since they allow understanding potential differences between the treatment groups in the time to onset of the first event. Since the pilot study involved three treatment groups – placebo, low dose, and high dose – it may be a good starting point to plot all groups first. Figure 3A shows a Kaplan-Meier plot with three strata corresponding to the treatments in the CDISC pilot study.

Even in the showcased scenario, we assume to have access to the clinical data, however, this might not be the case. Data protection is an important aspect of any research area. While data protection regulations have provided a way to share data and in return improve the reproducibility of experiments, in clinical research, sharing sensitive subject-specific data is impractical or simply not possible for legal reasons. Another option is to only share aggregated data or the analysis results. While this option can still bring privacy issues, for example due to the presence of outliers, results are already widely shared in publications through visualizations like the ones shown in Figs. 3 and 4. For Kaplan-Meier plots, this has led to numerous approaches<sup>17–20</sup> on extracting/retrieving the underlying results data, since these are often required e.g. in health technology assessments or when incorporating historical information into actual studies (e.g., Roychoudhury and Neuenschwander (2020)<sup>21</sup>). In contrast to current practice, having an ARDM in place gives many options on what data to share to support results reusability in a variety of contexts. For example, even regulatory agencies can benefit from the ARDM since outputs such as tables, graphics and listings can be easily generated from the results without the need to repeat or reproduce analyses. From our experience, it is common to initially share results with limited people (e.g., within a team) where we do not give much importance to details like aesthetics. However, at a later stage, researchers need the results to update the visualization to suit a wider audience, or use this data for future research. In the Kaplan-Meier plot example, this requires reverse-engineering by using tools to digitize the plot and create machine-readable results.

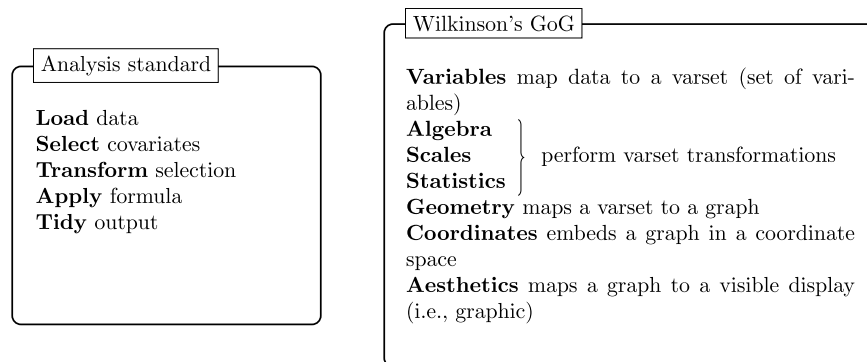


**Fig. 4** Employing an Analysis Results Data Model enables re-use at the results level rather than requiring source data. In this example, treatment arms can be removed (**A**), or additional summary statistics, such as the median survival time (**B**) or a risk table (**C**), can provide more context without repeating the underlying analysis.

A results visualization can appear in a variety of documents from presentation slides, an initial report, or a final publication, however, it is most likely not accompanied by the results used to create it. This hinders the reuse of the information (i.e., results) in the plot. A frequently encountered situation is illustrated in Fig. 4A, where one stratum is removed and the plot only shows two survival curves, for placebo and the high dose. This is not atypical in drug development, since after a general study overview, the focus is often on one dose only. While this update may seem trivial, from our experience, this task can require considerable time and effort due to the unavailability of the results. Without an analysis results data model or a known location where to find the results from the survival analysis, one must first locate the clinical data to perform the same analysis again. Then, search for and find the analysis code and the instructions to create the Kaplan-Meier plot. Eventually, one must repeat the analysis entirely. Thirdly, it is advisable to confirm whether the new plot matches the one we want to update; this is especially important if the analysis had to be redone as data transformations might have happened (e.g., different censoring than originally planned). Finally, one can filter the strata and create the plot in Fig. 4A.

## Methods

**The analysis results data model.** To create an analysis results data model, the first step requires thinking of the results of the analysis as data itself. Through this abstraction, we can begin organizing the data in a common model linking (e.g., clinical) datasets with the analysis results. Before we further introduce the ARDM it is necessary to clarify what an analysis and analysis results entail. An analysis is formally defined as a “detailed examination of the elements or structure of something”<sup>22</sup>. In practice, it is a collection of steps to inspect and understand data, explore a hypothesis, generate results, inferences, and possibly predictions. Analyses are fluid and can change depending on the conclusions drawn after each one of the steps. Nonetheless, routine analyses promote conventions that we can use as a foundation to create analysis standards. For example, looking at the



**Fig. 5** The analysis standard follows a grammar to define the steps in the analysis. Similarly, Wilkinson's<sup>23</sup> grammar of graphics (GoG) concisely defines the components required to produce a graphic.

table of contents of a Clinical Study Report (CSR) we can see a collection of routine results summaries. Diving deeper into these sections, we can see the same or similar analysis results between CSRs of independent clinical studies, namely due to conventions<sup>1</sup>. For example, it is standard for a clinical trial to report the demographics and baseline characteristics of the study population, and a summary of adverse events. These data summaries may also be a collection of separate data analyses grouped together in tables or figures (i.e., descriptive statistics of various baseline measurements, or the incidence rates of common adverse drug reactions, by assigned treatment). Also, the same *statistics*, such as the number of patients assigned to a treatment arm, may be repeated throughout the CSR. Complex inferential statistics may also be repeated in various tables and figures. For example, key outcomes maybe grouped together in a standalone summary of a drug's benefit-risk profile. Therefore, without upfront planning, the same *statistics* may be implemented many times in separate code.

The analysis results are the outcome of the analysis and are typically rendered into tables, figures, and listings to facilitate the presentation to stakeholders. Some examples of applications that can reuse the same results are present in Fig. 2 (right). Before the rendering, the results are stored in intermediate formats such as data frames or datasets. We can use this to our advantage and capture the results for posterior use in research by defining which elements to store and the respective constraints. This supports planning the analyses and the potential applications for the results, minimizing imprudent applications. An analysis results data model can be used to formalize the result elements to store and the constraints with the additional benefit of making the relationships between the results explicit. For example, we can store intermediate results, generated after the initial analysis steps, and use them to achieve the final analysis results. Besides improving the reusability of results, and reproducibility of the analysis, establishing relationships enables retracing the analysis steps and promotes transparency.

Data standards are useful to integrate and represent data correctly by specifying formats, units, and fields, among others. Due to the many requirements in clinical development, [guidelines detailing how to implement a data standard](#) are also frequent and essential to ensure the standard is correctly implemented and to describe the fundamental principles that apply to all data. An *analysis standard* would thus define the inputs and outputs of the analysis as well as the steps necessary to achieve those outputs. While an analysis convention follows a general set of context-dependent analysis steps, a standard ensures the analysis steps are inclusive (i.e., independent of context), consistent and uniform where each step is specified through a grammar<sup>23–25</sup> or the [querying syntax used in database systems](#). In Fig. 5, we compare the concepts behind an analysis standard with Wilkinson's grammar of graphics (GoG) data flow. Both follow an immutable order, ensuring that previous steps must be fulfilled to achieve the end result. For example, any data transformation needs to occur before we apply a formula (e.g., compute the descriptive statistics), otherwise, the result of the analysis becomes dubious. The collection of steps forms a grammar; however, each step also offers choices. For example, *apply formula* can refer to a linear model or Cox model. Wilkinson refers to this characteristic as the system's richness by the means of "paths" constructed by choosing different designs, scales, statistical methods, geometries, coordinate systems, and aesthetics. In the context of the ARDM, analysis standards support pre-planning, compelling the researcher to iterate over the potential analysis routes and the underlying question the analysis should address. In general, it is good practice to write down the details of an analysis, for example using a SAP, with sufficient granularity that the analysis could be reproduced independently if only the source data was available. Thus, the analysis standards would translate the intent expressed in the SAP into clear and well-defined steps.

Analysis standards bring immediate benefits to the analysis data quality<sup>26,27</sup> as it enables the validation of software and methods. With software validity, we refer to whether a piece of software does what it is expected and whether it clearly states how the output was reached. The validation of methods addresses whether the adequate statistical methodology was chosen. Due to its nature, this quality aspect is tightly related to other components of the clinical development process such as the SAP. In clinical development, standard operating procedures already cover many of these steps. However, they critically do not handle analysis results as a data source. Combining a data model with analysis standards would benefit clinical practice in four aspects:

1. Guaranteeing data quality and consistency across a clinical program, essentially creating a single source of

| Principle               | Justification  |
|-------------------------|--|
| Searchable              | • Consistent location to track and find results (i.e., knowledge source).  |
|                         | • Support search across results.   |
|                         | • Coherent querying through tools such as APIs.  |
|                         | • Accessible to data consumers and users.  |
| Nonredundant            | • Designed to acknowledge the <i>grammar</i> of data analysis (i.e., <i>estimate, store, retrieve, render</i> ). |
|                         | • Avoid repetitions (i.e., scripts and results).   |
|                         | • Control technical debt.  |
|                         | • Establish a single source of truth.  |
| Separation of concerns  | • Modularizes workflows to facilitate updates and expansion.   |
|                         | • Streamline validation and verification.  |
| Reusable and extensible | • Primary use: support ready-to-use outputs (e.g., plots for clinical reports).                                  |
|                         | • Secondary use: reusable for subsequent analyses (e.g., meta-analysis and systematic reviews).                  |
|                         | • Supports future extensions (e.g., novel outputs and updates to known applications).                            |
| Interoperable           | • Results and metadata are stored in a consistent data format.   |
|                         | • Results are represented by common or shared vocabularies (i.e., following a data model).                       |
|                         | • Enables standardization of outputs across programming languages bypassing language-specific conventions.       |
| Community-driven        | • Consensus building and accessible within and across organizations.   |
|                         | • Standards and tools are open, designed by and for the scientific community.                                    |

**Table 1.** The analysis results data model (ARDM) follows six principles broadly addressing the challenges in ensuring reproducible, traceable, reusable, and interoperable results.

truth designed to handle different levels of project abstraction. For example, from a single data analysis to a complete study, or a collection of studies.

2. Reusability by providing standardization across therapeutic areas and instigating the development of tools using the results instead of requiring individual patient data (e.g., interactive apps).
3. Simplicity as the analysis standard would encourage upfront planning and identify the necessary inputs, steps, and outputs to keep (e.g., reducing the complexity of forest plots and benefit-risk graphs summaries).
4. Efficiency by avoiding the manual and recurrent repetition of the analysis, and leveraging modularization and standardization of inferential statistics.

Analysis results datasets have been previously put forward as a solution to improve the uptake of graphics within Novartis, under the banner of graph-ready datasets<sup>28</sup>. Experienced study team leads often have implemented it for efficiency gains, especially around analysis outputs that would reuse existing summary statistics, for example, to support benefit-risk graphs where outcomes may come from different domains. Our experience has also revealed an element of institutional inertia. Standardizing analysis and results requires upfront planning which is often seen as added effort. However, teams that have gone through the steps of setting up a data model and a lightweight analysis process, have found efficiency and quality gains in reusing and maintaining code, as well as verifying and validating results. Regarding inferential results, instead of using results documents or repeating an analysis, we can simply access a common database where these are stored. An ARDM also simplifies modifications to the analysis (and consequently the results). With current practice, these changes might impact one function, program, or script in the best case, or multiple programs or scripts in the worst case. Using an ARDM only requires changes to one program as these can automatically propagate to any downstream analyses. The validation is also simplified as we transition from comparing data products (e.g., RTF files and plots) to comparing datasets directly. Additionally, this brings clarity and transparency, and is suitable for automation.

**Six guiding principles.** To create the ARDM we follow a collection of principles addressing the obstacles commonly faced during the clinical research process but also present in other areas. These principles are highlighted in Table 1 and broadly put forward improvements to quality, accessibility, efficiency, and reproducibility. On top of providing a data management solution, the ARDM compels us to take a holistic view of the clinical research process, from the initial data capture to the potential end applications. With this view, we have a clearer picture of where deficiencies occur and of their impact on the process.

The “searchable” principle refers to the easy retrieval of information by guaranteeing storage in a known, consistent, and technically-sound way. As we previously highlighted, it is common to have vast collections of results with very limited searchability. For example, figures in a collection of PDF documents. A practical solution is to have a data model to store the information consistently. In turn, this supports using a database that is by default more searchable than the PDF documents. With “searchable” in place, one can apply the “interoperable”, “non-redundant”, and “reusable and extensible” principles. In practice, this includes the use of consistent field names to store data in the database (e.g., the column “mean” has the mean value stored as a numeric value). The resulting coherent database is system-agnostic and can be queried through a variety of tools such as APIs. Thus, the data storing process supports straightforward querying which in turn can be used to avoid storing redundant results. Overall, this facilitates the use of the stored (results) data for primary analysis (i.e., submission to regulators) and secondary purposes (e.g., meta-analysis) but also allows for extensions of the data model granted the current

model constraints are respected. The “separation of concerns” refers to having the analysis (i.e., analysis code) separated from the source data, the results (e.g., from a survival analysis as shown in Fig. 3B), and the data products (e.g., the Kaplan-Meier plot in Fig. 3A). Finally, the “community-driven” principle ensures that the ARDM can be used pervasively, for instance, such that locations for tracking and finding results are not just multiplied across organizations but are community-developed and ideally lead to a single, widely accepted resource that can be searched as pioneered by the [EMBL GWAS Catalog](#).

In many industries where sub-optimal but quick solutions are preferred, [technical debt](#) is a growing problem. While some amount of technical debt is inevitable, understanding our processes can point us to where to make progressive updates and improvements. For example, upfront planning using analysis standards would reduce this debt by default as our starting point are previously verified and validated analyses (i.e., analysis standards). In an effort to continue reducing the debt, the ARDM separation of concerns principle streamlines changes and updates to processes since the analysis, results, and products are separate entities. Standardizing how to store results enables the use of different programming languages to perform analysis with traditionally non-comparable output formats (e.g., SAS and R). Furthermore, we believe the ARDM should grow organically and community-driven, supporting consensus building and cross-organization access.

## Discussion

The ARDM provides a solution to handle analysis results as data by creating a single source of truth. To guarantee the accuracy of the source, it leverages analysis standards (i.e., validated analyses) with known outputs which are then organized in a database following the proposed data model. The use of analysis standards supports the pre-planning of analyses, compelling the researcher to iterate on the best approach for analyzing the data, and potentially deciding to use pre-existing and appropriate analysis standards. Considering the ARDM from the biomedical data lifecycle view (e.g., through the lens of the [Harvard Medical School's Biomedical Data Lifecycle](#)), the ARDM touches the documentation & metadata, analysis ready datasets, data repositories, data sharing, and reproducibility stages. However, we take the point of view of a clinical researcher (both data consumer and producer) who sees the recurring problem of having to extract results data from published work. Therefore, in the context of the clinical trial lifecycle<sup>2</sup>, extending CDISC with the ARDM would touch on all of the biomedical data lifecycle phases as the ARDM relies on details present in supporting documents like the statistical analysis plan and data specifications.

The concept of creating standards through a common data model is recognised as good data management and stewardship practice. A few examples include the Observational Medical Outcomes Partnership data model, a standard designed to standardize the structure and content of observational data<sup>29</sup> and the Large-scale Evidence Generation and Evaluation across a Network of Databases research initiative to generate and store evidence from observational data<sup>30</sup>. The data model created by the [Sentinel initiative](#), led by the Food and Drug Administration (FDA), is tailored to organize medical billing information and electronic health records from a network of health care organizations. Similarly, the [National Patient-Centered Clinical Research Network](#) also established a standard to organize the data collected from their network of partners. Finally, expanding the search to translational medicine, the Informatics for Integrating Biology and the Bedside introduced a standard to organize electronic medical records and clinical research data<sup>31</sup>.

Alongside data models, standard processes have been established to generate analysis results such as the requirement to document analyses in SAPs<sup>32</sup>, including all data transformations from the source data to analysis ready data sets. However, analyses can be complex and dependent on technical factors, such as the statistical software used, as well as undocumented analysis choices throughout the pipeline, from source data to result. Even less complex routine analyses are error-prone and might not be clearly reproducible. Altogether, this process is time and resource-consuming. A proposed solution is to perform the analysis automatically. With this in mind and targeting clinical development, Brix *et al.*<sup>33</sup> introduce the ODM Data Analysis, a tool to automatically validate, monitor, and generate descriptive statistics from clinical data stored in the CDISC Operational Data Model format. The FDA's Sentinel Initiative is also capable of generating descriptive summaries and performing specific analysis leveraging the proprietary [Sentinel Routine Querying System](#).

Following this direction, the natural progression would be to create a standard suited for storing analysis results. Such an idea is implemented in the genome-wide association studies (GWAS) catalogue where curators assess GWAS literature, extract data, and store it following a standard including the [summary statistics](#). Taking a step in this direction, CDISC began the [360 initiative](#) to support the implementation of standards as linked metadata in an attempt to improve efficiency, consistency, and reusability across the clinical research. Nonetheless, the irreproducibility of research results remains an obstacle in clinical research and has brought up calls for global data standardization to enable semantic interoperability and adherence to the FAIR principles<sup>34</sup>. In our view, analysis standards and the ARDM are an important contribution to this initiative.

An important aspect which we did not explicitly discuss is the quality of the (raw/source) data which will ultimately serve as the source of any analyses for which results dataset are created through the ARDM. While the ARDM can be seen as a concept naturally tied to the CDISC philosophy, which is most prominently used in drug development studies that are conducted in a highly regulated environment with rigorous data quality standards, its applicability goes far beyond. For example, analyses conducted on open health data could also benefit from the ARDM, which would help to simplify traceability, exchangeability and reproducibility of analysis results. However, when working with these kind of data, understanding the quality of the underlying raw data is of paramount importance. In particular, since the ARDM will make analysis results more easily accessible and reusable also to an audience who may only have a limited understanding of how to assess the quality of the underlying raw data. In this wider context, it may be beneficial to use data quality evaluation approaches that were developed for a non-technical audience or for an audience without subject-matter (domain) expertise<sup>35</sup>. This will allow the audience to interpret the results taking the quality of the underlying raw data into account.



Utilizing the proposed ARDM has a set of requirements. Firstly, the provided clinical data must follow a consistent standard (i.e., CDISC ADaM). Our solution involves automatically populating a database, hence there are expectations regarding the structure of the data. Similarly, data standards are necessary to enable analysis standards. If the analysis input expectations are not met, the analysis is unsuccessful and no results are produced or stored. Further, when a data standard is updated it is necessary to also update the analysis standards and the ARDM accordingly. Another limitation is the necessity of analysis standards. Without quality analysis standards, the quality of the source of truth is not guaranteed. Creating analysis standards requires a good understanding of the analysis to correctly define the underlying grammar and identify relevant decision options for the user (e.g., filter data before modeling). The third limitation corresponds to the applications. At the moment, the ARDM stores and organizes results in a suitable way to reuse in known applications (e.g., creating plots, tables, and requesting individual result values). As future applications are unknown, the data model might not store all the information needed. However, given the ARDM modular approach, it is only necessary to update the result information to be kept rather than updating the entire workflow. Another limitation refers to the supported data modalities. The proposed ARDM is implemented on tabular clinical trial data. However, it is possible to adapt the ARDM and design choices (e.g., type of database) to support diverse data. For example, the [summary statistics](#) present in the genome-wide association studies (GWAS) catalog could be stored following an ARDM.

The current option to share and access clinical trials results is [ClinicalTrials.gov](#). Nonetheless, this is a repository and does not permit querying results as these are not stored as data (i.e., a machine-readable dataframe). The ARDM is an attempt to bring forward the problem of reproducibility and the lack of a single source of truth for analysis results. With it, we call for a paradigm shift where the target for the data analysis becomes the data model. Nonetheless, we understand the ARDM limitations and view it as one solution to a complex problem. We believe the best way to understand how the ARDM should evolve, or to shape it into a better solution, is to hear the opinions of the community. Hence, our underlying objective is to get the community's attention, discover similar initiatives, and converge on how to move forward in establishing analysis results as a data source to support future reusability and knowledge discovery.

### Data availability

The CDISC Pilot Project ADaM ADSL, ADTTE, and ADAE datasets were used to support the implementation of the analysis results data model. This data can be found at the PHUSE scripts repository ([https://github.com/phuse-org/phuse-scripts/blob/fa55614d7d178a193cc9b6e74256ea2d8dcf5d80/data/adam/TDF\\_ADaM\\_v1.0.zip](https://github.com/phuse-org/phuse-scripts/blob/fa55614d7d178a193cc9b6e74256ea2d8dcf5d80/data/adam/TDF_ADaM_v1.0.zip)) and at the repository supporting this manuscript<sup>15</sup>.

### Code availability

The implementation of the analysis results data model is available on Github<sup>15</sup>. This repository exemplifies how to construct the data model and the respective schema, as well as shows how to query the underlying database. Furthermore, we provide three output examples to visualize the results.

Received: 25 April 2022; Accepted: 18 October 2022;

Published online: 10 November 2022

### References

- European Medicines Agency. ICH Topic E 3 - Structure and Content of Clinical Study Reports. [https://www.ema.europa.eu/en/documents/scientific-guideline/ich-e-3-structure-content-clinical-study-reports-step-5\\_en.pdf](https://www.ema.europa.eu/en/documents/scientific-guideline/ich-e-3-structure-content-clinical-study-reports-step-5_en.pdf) (1996).
- Committee on Strategies for Responsible Sharing of Clinical Trial Data, Board on Health Sciences Policy & Institute of Medicine. Sharing clinical trial data (National Academies Press, Washington, D.C. 2015).
- Maciocci, Giuliano and Aufreiter, Michael and Bentley, Nokome. Introducing eLife's first computationally reproducible article. <https://elifesciences.org/labs/ad58f08d/introducing-elife-s-first-computationally-reproducible-article> (2019).
- R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/> (2021).
- Robinson, D., Hayes, A. & Couch, S. broom: Convert Statistical Objects into Tidy Tibbles. <https://CRAN.R-project.org/package=broom>. R package version 0.7.6 (2021).
- Siebert, M. *et al.* Data-sharing and re-analysis for main studies assessed by the european medicines agency—a crosssectional study on european public assessment reports. *BMC medicine* **20**, 1–14 (2022).
- Gelman, A. & Loken, E. The garden of forking paths: Why multiple comparisons can be a problem, even when there is no “fishing expedition” or “p-hacking” and the research hypothesis was posited ahead of time. *Dep. Stat. Columbia Univ.* **348** (2013).
- Wicherts, J. M. *et al.* Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Front. psychology* **1832** (2016).
- Devezer, B., Navarro, D. J., Vandekerckhove, J. & Ozge Buzbas, E. The case for formal methodology in scientific reform. *Royal Soc. open science* **8**, 200805 (2020).
- Higgins, J. P. *et al.* *Cochrane handbook for systematic reviews of interventions* (John Wiley & Sons, 2019).
- Tendal, B. *et al.* Disagreements in meta-analyses using outcomes measured on continuous or rating scales: observer agreement study. *BMJ* **339** (2009).
- Wilkinson, M. D. *et al.* The fair guiding principles for scientific data management and stewardship. *Sci. data* **3**, 1–9 (2016).
- Huser, V., Sastry, C., Breymaier, M., Idriss, A. & Cimino, J. J. Standardizing data exchange for clinical research protocols and case report forms: An assessment of the suitability of the Clinical Data Interchange Standards Consortium (CDISC) Operational Data Model (ODM). *J. biomedical informatics* **57**, 88–99 (2015).
- European Medicines Agency. European Medicines Regulatory Network Data Standardisation Strategy. [https://www.ema.europa.eu/en/documents/other/european-medicines-regulatory-network-data-standardisation-strategy\\_en.pdf](https://www.ema.europa.eu/en/documents/other/european-medicines-regulatory-network-data-standardisation-strategy_en.pdf) (2021).
- Barros, JM., A Widmer, L. & Baillie, M. Analysis Results Data Model, *Zenodo*, <https://doi.org/10.5281/zenodo.7163032> (2022).
- Kaplan, E. L. & Meier, P. Nonparametric estimation from incomplete observations. *J. Am. Stat. Assoc.* **53**, 457–481 (1958).
- Guyot, P., Ades, A., Ouwens, M. J. & Welton, N. J. Enhanced secondary analysis of survival data: reconstructing the data from published kaplan-meier survival curves. *BMC medical research methodology* **12**, 1–13 (2012).
- Liu, Z., Rich, B. & Hanley, J. A. Recovering the raw data behind a non-parametric survival curve. *Syst. reviews* **3**, 1–10 (2014).

19. Liu, N., Zhou, Y. & Lee, J. J. IPDfromKM: reconstruct individual patient data from published kaplan-meier survival curves. *BMC Med. Res. Methodol.* **21**, 1–22 (2021).
20. Rogula, B., Lozano-Ortega, G. & Johnston, K. M. A method for reconstructing individual patient data from kaplan-meier survival curves that incorporate marked censoring times. *MDM Policy & Pract.* **7** (2022).
21. Roychoudhury, S. & Neuenschwander, B. Bayesian leveraging of historical control data for a clinical trial with time-to-event endpoint. *Stat. medicine* **39**, 984–995 (2020).
22. Cambridge University Press. Analysis. In Cambridge Academic Content Dictionary, <https://dictionary.cambridge.org/dictionary/english/analysis>(Cambridge University Press, 2021).
23. Wilkinson, L. The grammar of graphics. In Handbook of computational statistics, 375–414 (Springer, 2012).
24. Wickham, H. Tidy data. *J. Stat. Softw.* **59**, 1–23 (2014).
25. Lee, S., Cook, D. & Lawrence, M. Plyranges: A grammar of genomic data transformation. *Genome biology* **20**, 1–10 (2019).
26. PhUSE Standard Analysis and Code Sharing Working Group. Best Practices for Quality Control and Validation. <https://phuse.s3.eu-central-1.amazonaws.com/Deliverables/Standard+Analyses+and+Code+Sharing/Best+Practices+for+Quality+Control+%26+Validation.pdf> (2020).
27. European Medicines Agency. ICH Topic E 6 - Guideline for good clinical practice (R2). [https://www.ema.europa.eu/en/documents/scientific-guideline/ich-e-6-r2-guideline-good-clinical-practice-step-5\\_en.pdf](https://www.ema.europa.eu/en/documents/scientific-guideline/ich-e-6-r2-guideline-good-clinical-practice-step-5_en.pdf) (2015).
28. Vandemeulebroecke, M. *et al.* How can we make better graphs? an initiative to increase the graphical expertise and productivity of quantitative scientists. *Pharm. Stat.* **18**, 106–114 (2019).
29. Observational Medical Outcomes Partnership. OMOP Common Data Model. <https://ohdsi.github.io/CommonDataModel/> (2021).
30. Schuemie, M. J. *et al.* Principles of large-scale evidence generation and evaluation across a network of databases (LEGEND). *J. Am. Med. Informatics Assoc.* **27**, 1331–1337 (2020).
31. Murphy, S. N. *et al.* Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). *J. Am. Med. Informatics Assoc.* **17**, 124–130 (2010).
32. Gamble, C. *et al.* Guidelines for the content of statistical analysis plans in clinical trials. *JAMA* **318**, 2337–2343 (2017).
33. Brix, T. J. *et al.* ODM data analysis—a tool for the automatic validation, monitoring and generation of generic descriptive statistics of patient data. *PLoS one* **13**, e0199242 (2018).
34. Jauregui, B. *et al.* The turning point for clinical research: Global data standardization. *J. Appl. Clin. Trials* (2019).
35. Nikiforova, A. Analysis of open health data quality using data object-driven approach to data quality evaluation: insights from a latvian context. In *IADIS International Conference e-Health*, 119–126 (2019).
36. Peter Van Reusel. CDISC 360: What's in It for Me? [www.cdisc.org/sites/default/files/2021-10/CDISC\\_360\\_2021\\_EU\\_Interchange.pdf](http://www.cdisc.org/sites/default/files/2021-10/CDISC_360_2021_EU_Interchange.pdf) (2021).

## Acknowledgements

We thank Carlotta Caroli, Nicholas Kelley, and Shahram Ebadollahi for their role in establishing and stewarding the AI4Life residency program. We also want to acknowledge Janice Branson for her valuable comments and support in this journey. Finally, J.M.B. would like to thank Idorsia Pharmaceuticals for the support during the final submission.

## Author contributions

All authors conceived and contributed to design the approach. J.M.B., L.A.W. and S.W. supervised the project. J.M.B. developed the data model and analysis standards. M.B. and L.A.W. reviewed the methodology. All authors read, edited, and approved the manuscript.

## Competing interests

The authors declare no competing interests.

## Additional information

**Correspondence** and requests for materials should be addressed to J.M.B. or M.B.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2022