UC San Diego
UC San Diego Previously Published Works

Title

Data Management 101 for drug developers: A peek behind the curtain.

Permalink

https://escholarship.org/uc/item/6s72t25w

Journal

Clinical and Translational Science, 16(9)

Authors

Oronsky, Bryan
Burbano, Erica
Stirn, Meaghan
et al.

Publication Date

2023-09-01

DOI

10.1111/cts.13582

Peer reviewed

Data Management 101 for drug developers: A peek behind the curtain

Bryan Oronsky[1] | Erica Burbano[1] | Meaghan Stirn[1] | Joanne Brechlin[1] |
Nacer Abrouk[2] | Scott Caroen[1] | Angelique Coyle[1] | Jeannie Williams[1] |
Pedro Cabrales[3] | Tony R. Reid[1]

[1]EpicentRx, Torrey Pines, California, USA

[2]Clinical Trial Innovations, Mountain View, California, USA

[3]UCSD Department of Engineering, La Jolla, California, USA

Correspondence
Bryan Oronsky, EpicentRx, Torrey Pines, CA, USA.
Email: boronsky@epicentrx.com

Abstract

In drug development a frequently used phrase is "data-driven". Just as high-test gas fuels a car, so drug development "runs on" high-quality data; hence, good data management practices, which involve case report form design, data entry, data capture, data validation, medical coding, database closure, and database locking, are critically important. This review covers the essentials of clinical data management (CDM) for the United States. It is intended to demystify CDM, which means nothing more esoteric than the collection, organization, maintenance, and analysis of data for clinical trials. The review is written with those who are new to drug development in mind and assumes only a passing familiarity with the terms and concepts that are introduced. However, its relevance may also extend to experienced professionals that feel the need to brush up on the basics. For added color and context, the review includes real-world examples with RRx-001, a new molecular entity in phase III and with fast-track status in head and neck cancer, and AdAPT-001, an oncolytic adenovirus armed with a transforming growth factor-beta (TGF-β) trap in a phase I/II clinical trial with which the authors, as employees of the biopharmaceutical company, EpicentRx, are closely involved. An alphabetized glossary of key terms and acronyms used throughout this review is also included for easy reference.

INTRODUCTION

To develop a drug is to undertake a journey, a long and costly one, full of twists, turns, and often unanticipated hazards. This long, costly, and complex journey starts with basic research and discovery, preclinical development tests, increasingly complicated human clinical trials, and ends, hopefully, with regulatory approval by the Food and Drug Administration (FDA).[1,2] The better the drug development "vehicle," the more likely it is to stay on track and to successfully navigate the almost always winding and rocky (as opposed to straight ahead and smooth) road to approval. Broadly, the components of this "vehicle" include the "engine" and the "chassis", that is, the drug development team, which undergirds, supports, and powers the vehicle, the wheels, or the clinical trial protocols, which advance it forward, and the "fuel" or data. Just as gas fuels a car, so drug development runs on data—and

not just any data but high-quality or "high octane," statistically interpretable data that supports the sponsor's labeling claims.

This overview, which is intended as a starting point for neophytes with little or no actual experience and as an aide-mémoire or checklist for the experienced individual, covers the essentials of clinical data management (CDM) for the United States both before and after the COVID-19 epidemic. It is written from the perspective of nine industry veterans with wide ranging drug development experience as healthcare providers (B.O., J.W., and T.R.R.), data managers (E.B. and J.B.), clinical research coordinators (M.S.), project managers and clinical research associates (CRAs) (S.C. and A.C.), and medical statisticians (N.A.). Surprisingly, given the pervasive nature of data and the critical importance of CDM to the success or failure of drug development, reviews or overviews on this topic are few and far between, and none are particularly recent or up to date, which was what motivated the writing of this review.

This term CDM is defined as the multistep process by which subject data are collected, protected, cleaned, and managed in compliance with a Code of Federal Regulations (CFR), 21 CFR Part 11, which applies to records in electronic form.[3] Since poor data quality,[4] of which error is a key determinant, undermines the confidence in and validity of clinical trial results and contributes to poor decision-making, all efforts must be undertaken to minimize error wherever and whenever possible. Real-world examples are provided with nibrozetone (RRx-001), a new molecular entity (NME) in phase III[5] that has received FDA Fast Track designation for the treatment of severe oral mucositis,[6] and AdAPT-001, an oncolytic adenovirus armed with a transforming growth factor-beta (TGF-β) trap in phase II[7,8] to add color and context since the authors are part of the drug development team that has successfully guided and shepherded these two entities from inception and preclinical testing to clinical trial evaluation in different indications from cancer to COVID-19.[9–11] An alphabetized glossary of key terms and acronyms (Table 1) used throughout this review is also included for quick reference in case these are unfamiliar to the reader.

CLINICAL DATA MANAGEMENT SYSTEMS (CDMS)

This term refers to the use of 21 Code of Federal Regulations (CFR) Part 11-compliant software, which applies to records in electronic records, to electronically store, capture, protect, and query data. CDMS are preferable to paper-based data capture because of their accuracy, consistency,

reliability, and auditability. However, universal adoption has not occurred largely because of cost, which is not easily borne by academic institutions or small biotechnology companies. Commercial systems like Oracle Clinical, InForm, Medrio, Macro, and Rave are often prohibitively expensive, requiring an investment of hundreds of thousands of dollars depending on the size of the trial and number of licenses needed. This is in comparison to open-source systems like DADOS Prospective, OpenClinica and TrialDB, which are more commonly used in an academic research setting.

In the first-in-man trial and for part of the first phase II trial called ROCKET for nibrozetone (RRx-001),[12–14] EpicentRx[15] used paper-based data capture (PDC) mainly due to two factors. The first was inconsistent internet connectivity, which is hugely ironic considering that the company was originally based in Silicon Valley, the birthplace of technology companies like Hewlett Packard, Intel, Apple, Cisco, and Google, and the second was the cost of around $150,000 for a commercial electronic database platform, which exceeded the allotted trial budget at the time. Ultimately, however, because of the impracticality and inefficiency of PDC, the company decided to make the switch first to Medrio, and then to Rave, despite the added expense involved.

Disadvantages of CDMS include the need for password assignment and regular password resets, high-speed network connectivity, secure data entry consoles and a web server, 24/7 support from a database manager, study-specific validation of the system, training of sponsor and clinical site personnel, and the occurrence of regular software upgrades, which may lead to data loss and considerable system downtime.[16]

CASE REPORT FORM (CRF)

The CRF is a customized and, hopefully, simple-to-use document, which is (or should be) designed to accurately capture relevant data and key variables and metrics of interest specific to a clinical trial protocol.[17] These data and metrics, which may include demographics, medical history, concomitant medications, dosing schedule of the test drug, clinical trial procedures, subject visits to the clinic, response rates, imaging scans such as CTs, MRIs, or X-rays, overall survival, and so on, vary from protocol to protocol depending on what indication and patient population are under study. All data collected on the CRF is de-identified. Figure 1 is an actual paper-based CRF form from a brain metastasis trial called BRAINSTORM[18] with RRx-001 that is labeled with partially redacted patient-specific information and provides instructions on how to fill it out.

TABLE 1 Glossary of key terms and acronyms.

Data management term	Definition
AdAPT-001	An engineered variant of the common cold virus equipped with a transforming growth factor-beta (TGF-β) "trap".
Adverse events (AEs)	Undesired effects of a drug that can range from mild to severe and can be life-threatening.
Blinding	Refers to the concealment or masking of group allocation from the subjects in a clinical trial, and possibly the investigators as well. In an open-label clinical trial no withholding of information from subjects or investigators occurs.
Case report form (CRF)	A printed, optical, or electronic document, which records all the information that is required by the clinical trial protocol.
Clinical Data Interchange Standards Consortium (CDISC)	Data standards developing organization for regulatory submissions including Standard for Exchange of Nonclinical Data (SEND) for nonclinical data, Study Data Tabulation Model (SDTM) for clinical data, and Analysis Data Model (ADaM) for analysis-ready data.
Clinical data management systems (CDMS)	Software tools available for CDM. Examples include Oracle Clinical, Clintrial, Rave, eClinical suite, and Macro.
Clinical outcome events (COEs)	These measure the result of a treatment or intervention. Typical examples of outcomes especially in cancer clinical trials are clinical worsening or progression, and mortality.
Clinical research associate (CRA)	A person that monitors the conduct of clinical trials on behalf of pharmaceutical companies.
Clinical study report (CSR)	An integrated report on a clinical trial presented in an easily searchable format in accordance with ICH E6 Section 1.13. CSRs are complete or abbreviated depending on whether the clinical trial is intended to support the efficacy claim for the dose, regimen, dose, regimen, population, or indication.
Clinical trial	Tests the safety and efficacy of new drugs or other interventions in preparation for an application to introduce them.
Code of Federal Regulations (CFR) Part 11	Refers to the FDA's regulations on electronic records and electronic signatures for clinical trials. Since electronic documents are in use everywhere, compliance with Part 11 is very important.
Database	A structured set of information, or data, typically stored and accessed electronically in a computer system.
Data capture	A process by which information is extracted from paper or electronic documents and converted into data for computer systems.
Data cleaning	The correction of errors and inconsistencies in a raw dataset in preparation for analysis.
Data entry	The transcription and input of data into an electronic format.
Data management plan (DMP)	A formal written document that describes how the data will be handled during and after a research project or a clinical trial.
Database lock	The step in a clinical trial when the database is locked or frozen to further modifications which include additions, deletions, or alterations of data in preparation for analysis. Also referred to as a "final lock" or "hard lock."
Electronic case report form (eCRF)	An auditable electronic record document, which records all the information that is required by the clinical trial protocol.
Electronic health record (EHR)	A digital record of health information.
Electronic source (eSource)	Initial electronic data capture.
Fast Track designation	An FDA program whose award expedites the review and development of drugs intended to address an unmet medical need.
Food and Drug Administration (FDA)	A United States government agency within the Department of Health and Human Services that oversees public health and approves or rejects human and veterinary drugs, vaccines, biological products, and medical devices for marketing.
Good clinical practice (GCP)	An international ethical and scientific quality standard that governs how to design, conduct, monitor, and report clinical trials.
HL7 Fast Healthcare Interoperability Resources (FHIR)	A set of rules and specifications for the exchange of electronic health care information.

(Continues)

TABLE 1 (Continued)

Data management term	Definition
Interim database lock	The process whereby part of a dataset is kept constant (i.e., frozen in time or locked) usually to perform an interim analysis of partial data for the purpose of making adjustments to the protocol or decisions about the clinical trial.
International Council for Harmonisation (ICH) E6	The guideline from the international non-profit organization called the ICH for good clinical practice (GCP).
Institutional Review Board (IRB)	Under FDA regulations, the IRB is an administrative committee that oversees clinical trials.
Medical coding	The categorization of medical terms and AEs for review and analysis.
Medical Dictionary for Regulatory Activities (MedDRA®)	A medical coding dictionary used by regulatory authorities and pharmaceutical companies to classify AEs.
Metadata	Data that describe information or context about other data such as the text or the image.
Protocol	A document colloquially referred to as "The Bible" that describes in depth exactly how the clinical trial will be conducted.
Randomization	Assignment or allocation of subjects (or patients) by chance to groups that receive different treatments. Most frequently, the investigational group receives the new treatment, and the control group receives standard therapy.
Regulatory submission	Any documentation or information submitted to a regulatory agency such as the FDA for review.
Remote monitoring	Off-site evaluation of a clinical site where the clinical trial is conducted that is performed by the clinical research associate (CRA).
Nibrozetone (RRx-001)	An experimental drug in late-stage clinical trials that targets an inflammatory complex called NLRP3 and that has received FDA Fast Track designation for severe oral mucositis.
Source data verification (SDV)	A process by which the information recorded in the CRF is compared with the original source records.
Sponsor	A company, institution, or organization that pays for and conducts a clinical trial.
Targeted source data verification (tSDV)	A process of monitoring that focuses only on critical data elements such as key eligibility and end point data rather than all data. The latter is referred to as 100% SDV.
Query	A request for information usually from a clinical site about data that are potentially inconsistent, incomplete, or missing.
Validation	A process to check if a set of data is rational and acceptable before its use.

FIGURE 1 Partially redacted study termination case report form from a brain metastasis clinical trial with RRx-001.[1,2]

CRF development, which requires specific expertise and knowledge, is indispensable to the success of a clinical trial as it represents the "output" of the text-based protocol in the form of data that statisticians can later restructure and summarize. A well-designed CRF should significantly reduce errors of commission and omission that might otherwise occur and, hence, the issuance of queries to correct those errors. Ideally, since the CRF derives from the protocol, its development should only commence after the protocol is finalized.

EpicentRx has followed suit with the current practice to replace traditional pen-and-paper collection forms[19] with electronic case report forms (e-CRFs), given that the latter is speedier, less onerous and cumbersome, more

environmentally friendly (because it largely eliminates paper), and less error-prone than the former. The adage "See one, (practice), do one, (practice), then teach one," which is a basic tenet of medical education,[20] applies also to CRF preparation.

DATA STANDARDIZATION

Central to data management is data standardization, the definition of which, as extracted from Richesson et al., refers to "consensual specifications for the representation of data from different sources or settings."[21] The purpose of data standardization in drug development is to develop a common language and reporting format for reporting, research, and analysis.

The US Food and Drug Administration (FDA) and the Japanese Pharmaceutical and Medical Devices Agency have adopted a set of global data standards for marketing authorization approval by a non-profit organization called the Clinical Data Interchange Standards Consortium (CDISC). The CDISC standards, which are globally recognized and widely used by the pharmaceutical industry, span the clinical research process to include the Standard for Exchange of Nonclinical Data (SEND), Protocol Representation Model (PRM), data collection for case report forms with Clinical Data Acquisition Standards Harmonization (CDASH), aggregation and tabulation with the Study Data Tabulation Model (SDTM), Analysis Data Model (ADaM), Questionnaires, Ratings, and Scales (QRS), and Operational Data Model (ODM) for exchange of data, as shown in Table 2.

The HL7 Fast Healthcare Interoperability Resources (FHIR) specification, which has emerged as the leading standard for the exchange of healthcare data, has the potential to complement the CDISC ODM standard, although to date no alignment of ODM and HL7 has occurred.[22]
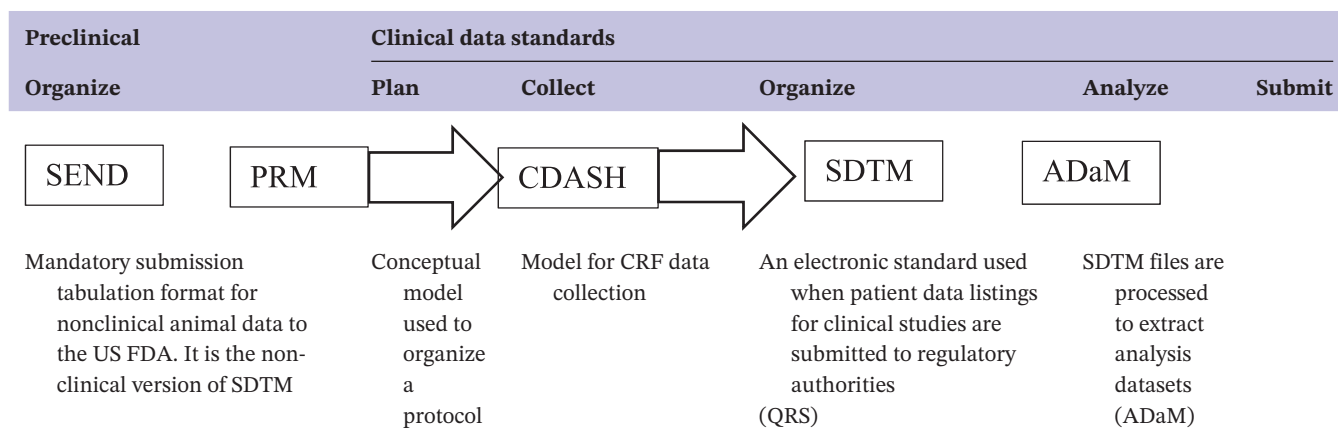
STORAGE

Data are currency and, as such, they can be bought and sold, or even stolen. Hence, all records and documents, and any data or information generated as part of a clinical trial, need to be securely stored and de-identified or anonymized for privacy. To de-identify data means to redact or to modify personal characteristics such as name, medical record number (MRN), date of birth/death, and so on.[23]

Hard copy information must be maintained in locked file cabinets in limited access, Institutional Review Board (IRB)-approved areas and electronic data must be protected from manipulation through unique access codes such as individual user ID and password combinations that are assigned only to those personnel with job responsibilities that require such access.[24] It is not safe to store or transport electronic data on unencrypted mobile devices such as laptops, tablets, smart phones, unencrypted external hard drives, or removable media like thumb drives, CDs, and DVDs.[25]

DATA MANAGEMENT PLAN (DMP)

As defined by the Society for Clinical Data Management (SCDM), a data management plan (DMP), which is protocol-specific, comprehensively documents data from their definition, collection, and processing to their final archival or disposal.[26] A DMP, which is usually reserved

TABLE 2 The Clinical Data Interchange Standards Consortium (CDISC) includes a collection of end-to-end standards that span the different stages in the clinical research process.

Preclinical	Clinical data standards				
Organize	Plan	Collect	Organize	Analyze	Submit
SEND	PRM	CDASH	SDTM	ADaM	
Mandatory submission tabulation format for nonclinical animal data to the US FDA. It is the non-clinical version of SDTM	Conceptual model used to organize a protocol	Model for CRF data collection	An electronic standard used when patient data listings for clinical studies are submitted to regulatory authorities (QRS)	SDTM files are processed to extract analysis datasets (ADaM)	

Abbreviations: ADaM, Analysis Data Model; CDASH, Clinical Data Acquisition Standards Harmonization; CRF, case report form; FDA, Food and Drug Administration; ODM, Operational Data Model for data exchange; PRM, Protocol Representation Model; QRS, Questions, Ratings, and Scales; SDM, Study Design Model; SDTM, Study Data Tabulation Model; SEND, Standard for Exchange of Nonclinical Data.

for larger clinical trials, covers source data verification, CRF design and completion guidelines, database design, procedures for data flow and data entry, data storage and protection, medical coding, data validation and query management, and database lock. It is an ongoing iterative document to be constantly refined and updated as necessary. The need to update the document is practically a given since clinical trial protocols frequently undergo amendments. The redacted table of contents from the DMP for the ongoing phase III trial, REPLATINUM, with RRx-001, which serves as a checklist for the required elements of a DMP, is shown in Figure 2.

SOURCE DATA VERIFICATION (SDV)

Data collection from clinical trials was and still is mainly a manual process. Back when EpicentRx used paper-based CRFs, clinical research coordinators (CRCs) at the clinical sites manually transcribed source data from patient charts to the CRFs.

EpicentRx clinical trial monitors (referred to as clinical research associates [CRAs]) regularly visited these sites; however, many there were for a given trial, and undertook either partial or 100% Source Data Verification (SDV). The purpose of SDV was to check that the data transcribed/reported recorded in the CRF matched the primary source data (e.g., patient medical records). In case of discrepancies, the CRAs queried the site staff and, if necessary, the site staff updated or amended the CRF. At this point, the paper CRFs were handed off to the CDM team. The CDM directly entered the CRFs into the database and issued further queries to address erroneous, missing, "out-of-range"/impossible, or inconsistent data (e.g., hysterectomy for a male patient).[27]

The nature and extent of the SDV was (and is) largely at the discretion of the sponsor since the FDA only recommends a review of a "representative" number of source documents rather than all of them.[28] Similarly, International Conference on Harmonisation (ICH) E6 recommendations state that "statistically controlled sampling may be an acceptable method for selecting the data to be verified," which suggests that 100% SDV is not necessary.[29,30]

To prevent any ambiguity relating to different interpretations of the term "source documents," the ICH Good Clinical Practice (GCP) requirements ("ICH E6 1.52") definition is included as follows: "All information in

F-DM-06-08
Protocol No.: RRx001-33

TABLE OF CONTENTS

FIGURE 2   REPLATINUM clinical trial data management plan table of contents.

original records and certified copies of original records of clinical findings, observations, or other activities in a clinical trial necessary for the reconstruction and evaluation of the trial. Source data are contained in source documents (original records or certified copies)."[31] The ICH E6 1.51 defines the term "source data" identically.

A prerequisite of SDV is to prove that the data in the CRFs are original, accurate, and verifiable through a traceable audit trail, as mandated by cGMP record-keeping practices; this audit trail, which documents the "who, what, when, where, why, and how" of any changes that were made to the data, applies to all records, irrespective of whether they are analog or digital.

The onset of COVID-19 lockdowns and suspension of in-person clinical site visits occasioned EpicentRx to permanently implement remote SDV via monitoring of electronic medical records. Here was a COVID silver lining and the very definition of a win–win, as it turned out; aside from the obvious convenience factor, remote monitoring hugely benefited the clinical sites, many of which were (and still are) too short-staffed and overburdened to support disruptive in-person visits, and the lack (or, more accurately, dearth) of travel saved EpicentRx a lot of time, money, and paperwork.

EpicentRx has since also moved away from 100% SDV, which was probably an example of overkill/ information overload, to targeted SDV (tSDV). With tSDV, the most important data such as inclusion/exclusion criteria, date of randomization, adverse events (AEs), concomitant medications, date of informed consent, protocol deviations, and so on are verified at random or "for cause" (i.e., only when a problem with one or more sites is identified). EpicentRx also remotely reviews investigator site files, site delegation logs, staff qualifications and training, and pharmacy documentation during tSDV.

In place of face-to-face meetings, which, admittedly, are the gold standard for establishing rapport and camaraderie,[32] EpicentRx has engaged with the clinical sites through alternative forms of communication in the form of telephone calls, text messaging, emailed training tips, short write-ups, and fun quizzes, as illustrated in Figure 3. The effect of this proactive information outreach strategy has been to encourage increased involvement and buy-in from the clinical sites, which, in turn, cuts down on the need for 100% SDV and nips (or, hopefully, nips) problems in the bud before they escalate.

DATA ENTRY

Data entry from paper-based forms is double or single. Double data entry is the definitive gold standard to identify and correct errors. Usually this involves two operators, each of whom enter the data separately and the datasets are compared for discrepancies. The resolution of these discrepancies is referred to as "verification." Single manual data entry, in which data are manually entered once, is also possible. Electronic CRFs (eCRFs) often eliminate one transcription step because CRCs fill them out, usually in real time, at the clinical sites.[3] Some eCRFs may also autopopulate from the electronic health record (EHR).

Sufficient time for data entry training of key personnel at the clinical site must be allocated, and logs should be kept as evidence of training. Immediate data entry is to be encouraged so that the most current, near real-time data are available for the sponsor to review and query. This makes possible immediate feedback to the clinical site for any issues in need of improvement. It is much more difficult to resolve data discrepancies and errors identified months after subjects have discontinued from the trial or clinical staff have moved on.

DATABASES AND EXTERNAL DATA

A clinical trial database incorporates all the data collected from the CRFs. A statistical database rearranges the data from the clinical trial database into a format for it to be statistically analyzed. The three major categories of statistical data are microdata or individual data, microdata or collective data, and metadata or data about data.

Not all the collected data come directly from the clinical trial itself. Some external data are imported from existing databases such as the electronic medical record (EMR), for example. However, the data extraction process from EMR systems, which occurs through manual abstraction, automated extraction, or a combination of both, is fraught with potential data errors, including mislabeling, subject identifier inconsistency, out-of-range values, duplications, and so on, which makes it important to conduct quality checks. Typically, this involves manual comparison between the manually or electronically extracted data and data direct from the source.

However, manual chart reviews are resource-intensive and difficult to scale with large clinical trials even if only a subset of the data is audited. Error prevention strategies with automated extraction include the mandatory completion of select data fields, dropdown data fields and autofill options, which leave little room for error, logic rules to constrain values and variables within certain ranges, accuracy rules (e.g., enrollment options for an adult cancer clinical trial do not include patients under the age of 18 years), and time-based rules (e.g., date of death cannot precede date of enrollment).[33]

Also, efforts to accelerate the adoption of electronic source (eSource) are underway, so that all source data, regardless of in what context they are acquired (e.g., office visit) and by whom (e.g., healthcare professional, patient, family member), are completely electronic, and fully acceptable for use in clinical trials and clinical trial submissions worldwide.[34]

DISCREPANCY MANAGEMENT

Discrepancy management refers to cleaning of data on the CRFs prior to its entry into the database, as defined earlier. The term is synonymous with query resolution.[3] A discrepancy is a query, or question, that flags any data in a CRF, which deviates from what is expected. Discrepancies include missing, "out-of-range," or inconsistent data (e.g., prostate cancer in a female or cervical cancer in a male). When discrepancies are flagged or identified, either from

automatic, built-in point-of-entry validation checks on the eCRFs or from manual validation checks of the eCRFs or paper CRFs, the clinical sites are asked for information or "clarification" (as shown in Figure 4) to resolve them and to sign and date the resolved documents, which provides an audit trail (i.e., reason for change, date change made, who made the change_. The faster the turnaround time on these queries, the sooner it is possible to start cleaning the data.

DATA CLEANING

The term "data cleaning" refers specifically to the identification and correction of errors in the data prior to analysis.[35] This straightforward definition aside, data cleaning often carries a negative connotation, as if it were a euphemism for post hoc data manipulation.[35] In fact, the whiff of stigma, which (unfairly) surrounds the neutral term, is so pervasive even statisticians as distinguished as Peter

Open Queries

Query ID: 80214831894   **Close Query**   **Assign**

Currently Assigned To: ████████████

Description: Please complete Death and Study Termination CRF

Designation: CRA

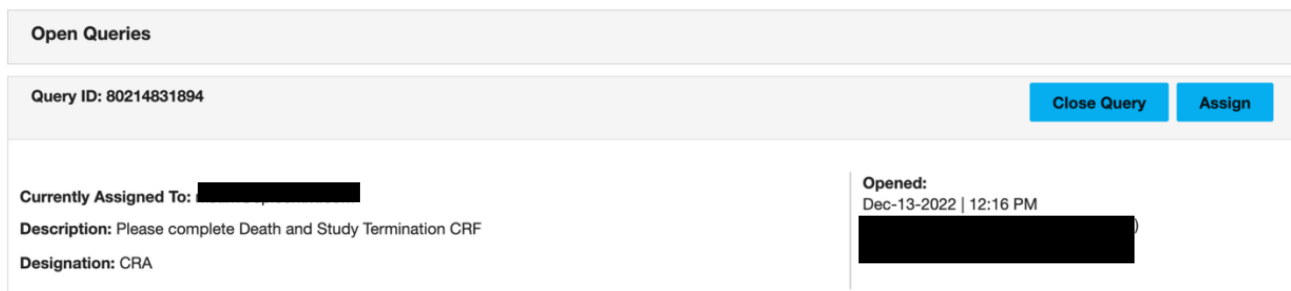Opened:
Dec-13-2022 | 12:16 PM

████████████████

FIGURE 4 Partially redacted query request for an RRx-001 clinical trial.

Armitage and Geoffrey Berry felt the need to semi-apologize for their inclusion of a chapter on data cleaning in their standard textbook on statistics in medical research.[36]

Nevertheless, data cleaning is essential to remove unwanted noise in the database.[37] As stated, this is a neutral term that refers to the detection and correction of duplicate, inconsistent, inaccurate, improperly formatted, or irrelevant data, all of which may reduce the reliability and robustness of the conclusions drawn. Its intent, therefore, is to put the data in the best possible shape for analysis, removing the "noise" but not the "signal," should one exist. To what degree cleaning is successful (i.e., to what degree it improves the signal-to-information ratio) depends on several factors including the formal, practical experience of the data managers and how well (or poorly) they work with big datasets, the data elements to be cleaned, the methods used to impute missing data points, and the robustness of the data management processes that are in place.

MEDICAL CODING

CRFs record unstructured information on adverse events (AEs), and concomitant medications (CMs). This information lacks uniformity as the non-standardized "verbatim" terms and descriptors for AEs and CMs are often replete with:

- Misspellings (e.g., "galebladder," "luopus," "acetomenophen," and "aspin")[26]
- Non-specific drug terms (e.g., "antibiotic")
- Medication brand names (e.g., Advil, Lipitor, and Vicoden)
- Truncations like "med" or "doc"
- Abbreviations like MI (myocardial infarction) and SOB (shortness of breath)[27]
- Idiomatically vague, imprecise, or incomplete expressions like "patient is circling the drain," "on the spectrum," "water in the lungs," or "hole in the heart."

Above average familiarity with clinical trials, industry acronyms, abbreviations, initialisms, informal jargon, and technical names of procedures and equipment is highly recommended for coders to convert the unstructured verbatim text in the CRF to standard, uniform medical terms so that healthcare providers, data managers, statisticians, and regulators will all be on the same page when it comes to review and analysis.[38–40] These uniform terms are associated with numeric/alphanumeric codes as identifiers.

The World Health Organization-Drug Dictionary Enhanced (WHO-DDE) is generally used to code medications.[41] The current standard for coding adverse drug reactions for the FDA, and other international regulatory entities per the ICH, is the Medical Dictionary for Regulatory Activities (MedDRA®). Under the oversight of the ICH, MedDRA is strictly maintained and regularly updated by the Maintenance and Support Services Organization (MSSO), which distributes licensed copies to users in industry and regulatory agencies.

The MedDRA taxonomy arranges codes like an upside-down ladder with five tiers or "rungs" from low to high specificity. At the top are 27 top-level hierarchies called System Organ Classes (SOCs). Below SOCs are high level group terms (HLGTs), high level terms (HLTs), preferred terms (PTs) and, finally, lower level terms (LLTs), as shown in Figure 5.

Coding is cognitively difficult since it depends on the translation of written, and not always easily decipherable, even with expert clinical knowledge, abbreviations and descriptions of diseases, diagnoses, and procedures. This is illustrated below with a real-life example taken from the phase II PREVLAR clinical trial in head and neck cancer,[6] where RRx-001 was administered as an anti-mucositis agent in combination with cisplatin and radiotherapy. A verbatim AE at one clinical site was "patient has blisters," an incomplete and misleading description, as it turned out, since the data management team sleuthed out that the blisters, in this case, were present not on the epidermis but in the oral mucosa. This was a key contextual piece of information that might have gone unnoticed and unquestioned, if not for the detective work of the data management team, whose intuition led them to dig deeper and to query the clinical site, since the incidence of severe oral
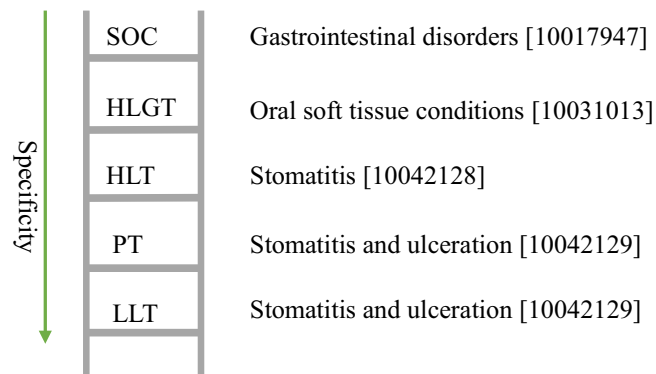
SOC	Gastrointestinal disorders [10017947]	
HLGT	Oral soft tissue conditions [10031013]	
HLT	Stomatitis [10042128]	
PT	Stomatitis and ulceration [10042129]	
LLT	Stomatitis and ulceration [10042129]	

Specificity

FIGURE 5 An actual example of Medical Dictionary for Regulatory Activities (MedDRA®) coding for "Patient Has Blisters," which referred to severe oral mucositis. HLGT, high level group term; HLT, high level term; LLT, lower level term; PT, preferred term; SOC, System Organ Class.

mucositis was a primary end point of the PREVLAR clinical trial.

A longstanding debate in coding is that between "lumpers" and "splitters." The former group "lump" together several AEs as manifestations of one condition, the latter separate seemingly related AEs into different categories. An actual example is an RRx-001-treated patient on the QUADRUPLE THREAT trial[42] with the following list of AEs, which occurred together: "influenza," "cough," "sore throat," "fever," "body aches," "fatigue," "coryza." A splitter codes each AE separately; a lumper subsumes them all under one entity, in this case influenza.

As a company, EpicentRx tends to err on the side of caution, which, on balance, probably leads it to split more than to lump. However, specific cases—and context—take precedence over general caution. When multiple symptoms, for example, coughing, sneezing, body aches, runny nose, and fever are present and these are clearly related to, and subsumed by, a single pathophysiologic entity, influenza, then from the authors' perspective it makes sense to lump them together and not to split them apart. Conversely, when the presence of signs and/or symptoms is simply suggestive, rather than definitive, of a syndromic etiology, then the authors prefer to split them up instead of grouping them together. In short, EpicentRx makes decisions not in a vacuum but on a case-by-case basis, considering the specifics and the context of the clinical data, the patients, and the information source.

DATABASE LOCK

The culmination of the preceding steps to collect, verify, validate, and clean the data is a final or "hard" database lock. A "hard locked" database is ready for analysis, and no further changes are expected or permitted. The time window between availability of the data and database lock varies from clinical trial to clinical trial and may take up to several months.[43]

A "soft" lock during which a final comprehensive review of the data is undertaken usually precedes a hard lock. An interim database lock is also performed in some clinical trials. Its purpose is to take a static "snapshot in time" of current data at a prospectively determined date (e.g., 52 weeks) or on a milestone basis (e.g., when 25% or 75% of patients are enrolled), at which point the blind is broken, if one is in place, usually for assessment and reporting purposes. However, an interim lock does not imply that data are in their final state and multiple interim locks are possible. During an interim lock, unblinding usually occurs for the sponsor, but not for the investigator, participants, and study-site personnel. Note that the data from an interim lock may lead to early termination of the trial if the experimental treatment is substantially more effective, or substantially less effective than the level the trial was designed to detect. The term "unblinding" refers to the disclosure of treatment group assignment during the trial, which was put in place to minimize conscious or unconscious biases that might come from knowledge of treatment.

CONCLUSIONS

The aim of this review was to present a holistic overview of clinical data management based on the perspective and reflections of an experienced team of drug development professionals with seven drug and device approvals between them. These several approvals aside, which are beyond rare, drug development is an expensive, lottery-like endeavor that most often comes to naught, given a failure/attrition rate of over 90%.[44] In other words, like a high stakes poker game, the entry fees are high, and the probability of a win/payout is low.[45] Even if a successful outcome is deemed to be probable, it is never guaranteed, given all the organizational, financial, regulatory, and clinical obstacles to overcome, and their associated moving parts, which stack the deck firmly against, rather than in favor of, and this is especially the case for new chemical and biological entities like nibrozetone (RRx-001)[46–49] and AdAPT-001[50] where the road ahead is invariably rocky. To beat these formidable odds in pursuit of an approval requires the proper generation, acquisition, maintenance, and use of high-quality data, which underpins not only drug approval, but also decision-making, labeling, and marketing.

The objective of data management is to catch errors at the earliest possible point or, ideally, to prevent them entirely. In the authors' experience, all it takes is some

missing or erroneous data (i.e., "garbage in") to impact the validity of the trial (i.e., "garbage out"). Nevertheless, because to err is human, datasets, which may contain billions of bytes of information, are never entirely error free or "clean", and no matter the data management processes and safeguards that are put in place to prevent errors, they still manage to creep in and, even worse, to propagate. Ultimately, the effectiveness of these processes and safeguards is only as good as their implementation. The more that personnel understand the specifics of the clinical trial protocol, and not only what errors to look for, but also which ones are the most relevant and important (because of their relationship to the primary and secondary end points of the trial, for example), the more likely these errors are to be discovered and corrected. Clearly, not all errors are created equal—some, like misspellings, typos, and out-of-range dates, are minor, and some like randomization of ineligible patients or noncompliance with important protocol procedures, for example, are major because they drastically impact the conduct and the interpretation of the clinical trial.

This perhaps goes without saying, but practice—and practical experience—in data management, as elsewhere, makes (almost) perfect. From the personal perspective of the authors, it is only with repeated exposure that "common sense" knowledge and skilled pattern recognition develop sufficiently to "diagnose" potentially harmful errors, ambiguities, and inconsistencies in the data, which may need to be flagged for further review. This "sixth sense" or "gut instinct" led the data management team to flag the term "blisters" (mentioned in the previous Medical coding section) for query, since in first-line head and neck cancer blisters may develop in the mouth. The point is that it takes time to hone instinct and intuition, the better to sort the wheat from the chaff, especially when the volume of data is large.

For clarity, the difference between a data management review and a medical review of the data is that the former concerns itself with the consistency and accuracy of the information that has been collected across all participants and sites, while the latter generally involves close attention to (1) predefined clinical outcome events (COEs), for example, all-cause mortality; (2) AEs, which represent untoward medical occurrences that are causally related or not with the investigational product; and (3) laboratory values, clinical events, and patient-reported outcome measures.

It seems fitting to penultimately conclude with the following prescriptive phrase in data management, which one of the reviewers of this article suggested that we include, namely, "begin with the end in mind." This phrase from Stephen Covey[51] refers to the backward design of protocols, CRFs, and the database in anticipation of the content and format of the final clinical study report (CSR); the CSR integrates the full efficacy and safety data complete with tables and figures for an individual study with a therapeutic or diagnostic agent. Depending on several factors including whether the study is uncontrolled or controlled, or whether it evaluates conditions related or unrelated to those for which a claim is/will be made, determines whether the CSR is to be written in full or abbreviated with summarized data and with some sections deleted.[52]

However, to "begin with the end in mind," while useful as a rule of thumb, comes with a caveat: clinical research is unpredictable and ever-changing. A priori premises and rationales are not static or absolute but remain fluid, potentially rendering them less valid or even invalid as time goes on. In this case, it is important to remain flexible enough to rework or scrap the original plan in accordance with the data obtained and to potentially restart with another end in mind.

As a final comment, a well-worn maxim, which has been modified herein, states that "behind every successful (wo)man there is a supportive spouse." This also applies to all FDA-approved drugs, which are inextricably married, usually for better and not for worse, to the quality and the integrity of the clinical trial data that support them.

CONFLICT OF INTEREST STATEMENT
B.O., E.B., M.S., J.B., S.C., A.C., J.W., and T.R.R. declare that they are employed by EpicentRx, the company that is referred to in this review. All other authors declared no competing interests for this work.

ORCID
Bryan Oronsky https://orcid.org/0000-0002-3940-6329

REFERENCES
1. Institute of Medicine (US) Committee on Conflict of Interest in Medical Research, Education, and Practice. The pathway from idea to regulatory approval: examples for drug development. In: Lo B, Field MJ, eds. *Conflict of Interest in Medical Research, Education, and Practice.* National Academies Press; 2009:373-385.
2. Oronsky B, Caroen S, Brinkhaus F, Reid T, Stirn M, Kumar R. Patent and marketing exclusivities 101 for drug developers. *Recent Pat Biotechnol.* 2023;17:257-270. doi:10.2174/1872208317666230111105223

3. Krishnankutty B, Bellary S, Kumar NB, Moodahadu LS. Data management in clinical research: an overview. *Indian J Pharmacol.* 2012;44(2):168-172. doi:10.4103/0253-7613.93842

4. Kahn MG, Brown JS, Chun AT, et al. Transparent reporting of data quality in distributed data networks. *EGEMS (Wash DC).* 2015;3(1):1052. doi:10.13063/2327-9214.1052

5. Oronsky B, Reid TR, Larson C, et al. REPLATINUM phase III randomized study: RRx-001 + platinum doublet versus platinum doublet in third-line small cell lung cancer. *Future Oncol.* 2019;15(30):3427-3433. doi:10.2217/fon-2019-0317

6. Bonomi M, Blakaj DM, Kabarriti R, et al. PREVLAR: phase 2a randomized trial to assess the safety and efficacy of RRx-001 in the attenuation of oral mucositis in patients receiving head and neck chemoradiotherapy. *Int J Radiat Oncol Biol Phys.* 2023;116(3):551-559. doi:10.1016/j.ijrobp.2022.12.031

7. Oronsky B, Gastman B, Conley AP, Reid C, Caroen S, Reid T. Oncolytic adenoviruses: the cold war against cancer finally turns hot. *Cancers (Basel).* 2022;14(19):4701. doi:10.3390/cancers14194701

8. Larson C, Oronsky B, Reid T. AdAPT-001, an oncolytic adenovirus armed with a TGF-β trap, overcomes in vivo resistance to PD-L1-immunotherapy. *Am J Cancer Res.* 2022;12(7):3141-3147.

9. Oronsky B, Takahashi L, Gordon R, Cabrales P, Caroen S, Reid T. RRx-001: a chimeric triple action NLRP3 inhibitor, Nrf2 inducer, and nitric oxide superagonist. *Front Oncol.* 2023;13:1204143. doi:10.3389/fonc.2023.1204143

10. Hammond TC, Lee RC, Oronsky B, et al. Clinical course of two patients with COVID-19 respiratory failure after administration of the anticancer small molecule, RRx-001. *Int Med Case Rep J.* 2022;15:735-738.

11. Oronsky B, Knox S, Cabrales P, Oronsky A, Reid TR. Desperate times, desperate measures: the case for RRx-001 in the treatment of COVID-19. *Semin Oncol.* 2020;47(5):305-308. doi:10.1053/j.seminoncol.2020.07.002

12. Reid TR, Abrouk N, Caroen S, et al. ROCKET: phase II randomized, active-controlled, multicenter trial to assess the safety and efficacy of RRx-001 + irinotecan vs. single-agent regorafenib in third/fourth line colorectal cancer. *Clin Colorectal Cancer.* 2023;22(1):92-99. doi:10.1016/j.clcc.2022.11.003

13. Reid T, Oronsky B, Scicinski J, Scribner CL, et al. Safety and activity of RRx-001 in patients with advanced cancer: a first-in-human, open-label, dose-escalation phase 1 study. *Lancet Oncol.* 2015;16(9):1133-1142. doi:10.1016/S1470-2045(15)00089-3

14. Jayabalan N, Oronsky B, Cabrales P, et al. A review of RRx-001: a late-stage multi-indication inhibitor of NLRP3 activation and chronic inflammation. *Drugs.* 2023;83(5):389-402. doi:10.1007/s40265-023-01838-z

15. Oronsky B. Profile EpicentRx. *Inc Hum Vaccin Immunother.* 2023;19(1):2184963. doi:10.1080/21645515.2023.2184963

16. Shah J, Rajgor D, Pradhan S, McCready M, Zaveri A, Pietrobon R. Electronic data capture for registries and clinical trials in orthopaedic surgery: open source versus commercial systems. *Clin Orthop Relat Res.* 2010;468(10):2664-2671. doi:10.1007/s11999-010-1469-3

17. Nahm M, Shepherd J, Buzenberg A, et al. Design and implementation of an institutional case report form library. *Clin Trials.* 2011;8(1):94-102. doi:10.1177/1740774510391916

18. Kim MM, Parmar HA, Schipper M, et al. BRAINSTORM: a multi-institutional phase 1/2 study of RRx-001 in combination with whole brain radiation therapy for patients with brain metastases. *Int J Radiat Oncol Biol Phys.* 2020;107(3):478-486. doi:10.1016/j.ijrobp.2020.02.639

19. Ene-Iordache B, Carminati S, Antiga L, et al. Developing regulatory-compliant electronic case report forms for clinical trials: experience with the demand trial. *J Am Med Inform Assoc.* 2009;16(3):404-408. doi:10.1197/jamia.M2787

20. Ayub SM. "See one, do one, teach one": balancing patient care and surgical training in an emergency trauma department. *J Glob Health.* 2022;12:3051. doi:10.7189/jogh.12.03051

21. Richesson RL, Krischer J. Data standards in clinical research: gaps, overlaps, challenges and future directions. *J Am Med Inform Assoc.* 2007;14(6):687-6s96. doi:10.1197/jamia.M2470

22. Vorisek CN, Lehne M, Klopfenstein SAI, et al. Fast health-care interoperability resources (FHIR) for interoperability in health research: systematic review. *JMIR Med Inform.* 2022;10(7):e35724. doi:10.2196/35724

23. Committee on Strategies for Responsible Sharing of Clinical Trial Data; Board on Health Sciences Policy; Institute of Medicine. Sharing clinical trial data: maximizing benefits, minimizing risk. *Appendix B. Concepts and Methods for De-identifying Clinical Trial Data.* National Academies Press; 2015:793-794.

24. Gliklich RE, Dreyer NA, Leavy MB, eds. Data collection and quality assurance. *Registries for Evaluating Patient Outcomes: A User's Guide.* 3rd ed. Agency for Healthcare Research and Quality; 2014:251-276.

25. Dhudasia MB, Grundmeier RW, Mukhopadhyay S. Essentials of data management: an overview. *Pediatr Res.* 2023;93(1):2-3. doi:10.1038/s41390-021-01389-7

26. Williams M, Bagwell J, Nahm ZM. Data management plans: the missing perspective. *J Biomed Inform.* 2017;71:130-142. doi:10.1016/j.jbi.2017.05.004

27. Gliklich RE, Leavy MB, Dreyer NA, eds. Obtaining data and quality assurance. *Registries for Evaluating Patient Outcomes: A User's Guide.* 4th ed. Agency for Healthcare Research and Quality; 2020:126-127.

28. Getz KA. Low hanging fruit in the fight against inefficiency. *Appl Clin Trials.* 2011;20:30-32.

29. International Council for Harmonisation Good Clinical Practice (ICH-GCP) – E6 (R1). 1996. Accessed April 20, 2023. https://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E6/E6_R1_Guideline.pdf

30. Smith PG, Morrow RH, Ross DA. Data management. *Field Trials of Health Interventions: A Toolbox.* 3rd ed. Oxford: Oxford University Press; 2015:145-158.

31. Bargaje C. Good documentation practice in clinical research. *Perspect Clin Res.* 2011;2(2):59-63. doi:10.4103/2229-3485.80368

32. Fougerou-Leurent C, Laviolle B, Bellissant E. Cost-effectiveness of full versus targeted monitoring of randomized controlled trials. *Fundam Clin Pharmacol.* 2018;32(S1):49.

33. Yin AL, Guo WL, Sholle ET, et al. Comparing automated vs. manual data collection for COVID-specific medications from electronic health records. *Int J Med Inform.* 2022;157:104622. doi:10.1016/j.ijmedinf.2021.104622

34. Parab AA, Mehta P, Vattikola A, et al. Accelerating the adoption of eSource in clinical research: a TransCelerate point of view. *Ther Innov Regul Sci.* 2020;54(5):1141-1151. doi:10.1007/s43441-020-00138-y

35. Van den Broeck J, Cunningham SA, Eeckels R, Herbst K. Data cleaning: detecting, diagnosing, and editing data abnormalities. *PLoS Med.* 2005;2(10):e267. doi:10.1371/journal.pmed.0020267

36. Armitage P, Berry G. *Statistical Methods in Medical Research*. 2nd ed. Blackwell Scientific Publications; 1987:559.

37. Houston L, Probst Y, Yu P, Martin A. Exploring data quality management within clinical trials. *Appl Clin Inform*. 2018;9(1):72-81. doi:10.1055/s-0037-1621702

38. Jiang K, Chen T, Huang L, Calix RA, Bernard GR. A data-driven method of discovering misspellings of medication names on twitter. *Stud Health Technol Inform*. 2018;247:136-140.

39. Tse T, Soergel D. Exploring medical expressions used by consumers and the media: an emerging view of consumer health vocabularies. *AMIA Annu Symp Proc*. 2003;2003:674-678.

40. Babre D. Medical coding in clinical trials. Perspect. *Clin Res*. 2010;1(1):29-32.

41. Uppsala Monitoring Centre. WHODrug Global. [cited 2020 February 2023]. https://www.who-umc.org/whodrug/whodrug-portfolio/whodrug-global/

42. Lee MJ, Tomita Y, Yuno A, et al. Results from a biomarker study to accompany a phase II trial of RRx-001 with reintroduced platinum-based chemotherapy in relapsed small cell carcinoma. *Expert Opin Investig Drugs*. 2021;30(2):177-183. doi:10.1080/13543784.2021.1863947

43. Ganju J. Improving the operational efficiency of phase 2 and 3 trials. *Trials*. 2016;17(1):332. doi:10.1186/s13063-016-1465-3

44. Kola I, Landis J. Can the pharmaceutical industry reduce attrition rates? *Nat Rev Drug Discov*. 2004;3(8):711-715. doi:10.1038/nrd1470

45. Wishart DS. Improving early drug discovery through ADME modelling: an overview. *Drugs R D*. 2007;8(6):349-362. doi:10.2165/00126839-200708060-00003

46. Reid T, Oronsky B, Caroen S, et al. Phase 1 pilot study of RRx-001 + nivolumab in patients with advanced metastatic cancer (PRIMETIME). *Front Immunol*. 2023;14:1104753. doi:10.3389/fimmu.2023.1104753

47. Caroen S, Oronsky B, Reid T, Pandher K, Lopez A. Superficial venous-associated inflammation from direct IV administration of RRx-001 in rats. *Int J Med Sci*. 2022;19(11):1628-1630. doi:10.7150/ijms.76615

48. Oronsky B, Caroen S, Abrouk N, Reid TR. RRx-001 and the "right stuff": protection and treatment in outer space. *Life Sci Space Res (Amst)*. 2022;35:69-75. doi:10.1016/j.lssr.2022.05.001

49. Kanter J, Oronsky B, Reid T, et al. Explosive hazards identified during the manufacture and transportation of 1-bromoacetyl-3,3-dinitroazetidine (RRx-001). *Organic Process Res Dev*. 2022;26(11):3010-3014. doi:10.1021/acs.oprd.2c00109

50. Larson C, Oronsky B, Abrouk NE, Oronsky A, Reid TR. Toxicology and biodistribution of AdAPT-001, a replication-competent type 5 adenovirus with a trap for the immunosuppressive cytokine, TGF-beta. *Am J Cancer Res*. 2021;11(10):5184-5189.

51. Covey SR. *The Seven Habits of Highly Effective People: Restoring the Character Ethic*. New York: Simon and Schuster; 1989.

52. CPMP/ICH. *137/95. ICH Topic E3. Structure and Content of Clinical Study Reports*. European Medicines Agency; 1996.