AAPS Open

PERSPECTIVE

Open Access

A controlled vocabulary and taxonomy for the submission of quality attributes for therapeutic proteins

Joel T. Welch[1*], Steven Kozlowski[1], Bazarragchaa Damdinsuren[1] and Brian A. Roelofs[1]

Abstract

Structured product quality data offer tremendous promise to revolutionize the submission of drug applications. However, the quality attributes for biological products do not have a systematic naming taxonomy, and consequently this limit poses a critical challenge in the development of systems for structured regulatory submissions. Here, we describe the creation of a controlled vocabulary with a structured taxonomical naming approach for quality attributes of therapeutic proteins. Additionally, we endeavor to make the case for why such systematic harmonized naming is required to support the successful implementation of structured data systems. We also describe the key principles of our structured naming approach, including a top-down view of the product and protein structure and a distinction between a quality attribute and the test to evaluate the attribute. Finally, we describe how this approach can accommodate emerging product types, advanced manufacturing technologies, and be used across the variety of submission sections in a regulatory dossier that discusses quality attributes.

Keywords Drug application, Biopharmaceutical characterization, Biosimilar, Biotherapeutic, Protein structure, Quality attribute, Protein, Structure, Structured data

Introduction

The British mathematician George Box is credited with the observation: “All models are wrong; some are useful”. This famous and highly quotable line has been used repeatedly to acknowledge a variety of truths about models and the systems they govern, namely: 1) They have fundamental limitations of extrapolation, 2) That they may be constructed from assumptions that might not be universally true for all cases, and 3) They may represent or predict an outcome that may not be consistent with absolute truth of the “real” result. Herein we describe the creation of a “model” controlled vocabulary and taxonomy for protein quality attributes classification and do so with an a priori acknowledgement that any such naming approach will have within it, limitations and counterintuitions. Nevertheless, we propose such a “model” that may provide a useful starting point for discussion and attempt to make a case for why such a system (or any universally agreed upon system) is desperately needed to support the successful implementation of structured data systems for biological products. The framework described herein is designed to apply to therapeutic proteins as other modalities (e.g., cell and gene therapies and vaccines) would potentially have other considerations.

Many recent publications have focused on the opportunities that progress in data storage, data structure, and data engineering offer the pharmaceutical industry (Algorri et al. 2020, 2022; Macdonald et al. 2021; Robertson et al. 2019). Much has been made of the opportunities

This publication reflects the views of the authors and should not be construed to represent FDA’s views or policies.

*Correspondence:
Joel T. Welch
Joel.Welch@fda.hhs.gov
[1] Center for Drug Evaluation and Research, Office of Pharmaceutical Quality, Office of Product Quality Assessment III, U.S. Food and Drug Administration, Silver Spring, USA

such advancements in technology offer to modernize the regulatory submission process, to facilitate a global, simultaneous submission, and with it, concurrent global assessment by different regulatory agencies. This opportunity is particularly appealing for pharmaceutical companies which generate tremendous volumes of data and information that ultimately require manual construction and collation prior to regulatory submission completion (Ahluwalia et al. 2022; Cauchon et al. 2019). The manual curation of these data does not occur as a single event during product development, rather they extend across a product lifecycle and across different regional geographies, and are one reason why global harmonization remains such a challenging yet appealing proposition. Furthermore, the systems used to house these data may comprise a "data lake" ecosystem where much of the data themselves may reside in a loose tapestry of a variety of systems (Beierle et al. 2023).

Regulatory landscape

Pharmaceutical manufacturing is itself undergoing a digital transformation, one that is frequently referred to as Pharma 4.0, or Industry 4.0 (Arden et al. 2021). This term refers to a fourth revolution, one that speaks to more than just the *digitizing* and the collection of physical data and their translation into an electronic format, but also the *digitalization* which is the conversion of a once human process into a computer-operated or automated one. This includes leveraging tools and technology such as automatization/robotics, artificial intelligence/machine learning, and advanced computing.

Paralleling the revolution underway in industry, regulatory health authorities also have tremendous interest and efforts in such transformations. One such prominent effort is FDA's Knowledge-Aided Assessment and Structured Application (KASA) initiative (Rosencrance et al. 2019). This tool, initially announced in 2018 and developed by the Office of Pharmaceutical Quality, is designed to use structured approaches for assessment to allow for more consistency, reproducibility and searchability (Brennan 2018). While developed initially with abbreviated new drug application (ANDA) submissions in mind, subsequent developments and discussion of the program have highlighted the potential and intent to use the program across all assessment programs, including new drug applications (NDAs) and biologics license applications (BLAs) (Cox 2021). Notably, while the "SA" of KASA stands for *structured application*, other related and interconnected initiatives (rather than KASA) may serve as the intended entry mechanism for approaching the structure of the data itself in the regulatory submission, as the KASA system currently serves as an assessment tool to be used internally by FDA. To that end, FDA announced the availability of draft documents to begin the conversation for standardizing pharmaceutical quality/chemistry manufacturing controls data (known as PQ/CMC), including a publicly available roadmap in 2017 (FDA 2019). The intent of the program is the elimination of data and information being submitted in PDF style format, and rather providing defined elements using data in a prescribed computer "understandable" language. This program included a pilot project in 2020 that utilized Health Level Seven International's (HL7) Fast Healthcare Interoperability Resources (FHIR) as their backbone (Schmuff 2019). The building blocks themselves reflect prespecified descriptive elements, such as for a batch: its batch number, product name, strength, batch size, manufacturing date, etc. A PQ/CMC draft published in 2022 was a critical first step and focused on specific data rich submission topics for oral solid drug presentations, such as Specification, Batch Analysis and Stability (FDA 2023).

Structured data approaches and interest are not limited to FDA, indeed other health authorities, in particular the European Medicines Agency (EMA) are supportive and developing such electronic formats. Most notably, efforts include the IDMP (Identification of Medicinal Products International Organisation for Standardisation,) which also utilizes HL7 for product information (EMA 2021). The IDMP reflects a suite of standards, including Medicinal Product Identification (MPID), Pharmaceutical Product Identifier (PhPID), Substance Identification (SubID), Dosage Form and Route of Administration, and Units of Measurement (UoM). Most relevant to our proposal here is the SubID, although the technical work instructions for it do not provide comprehensive controlled vocabularies for protein product quality attributes. A variety of other organizations have also proposed standardization approaches, though none have yet included a comprehensive proposal for quality attributes of proteins (Allotrope Foundation n.d.; Pistoia Alliance n.d.). Certain efforts to create a framework for other data elements are also well underway. For example, FDA has adopted certain standards, in particular ISO 11238 standard to aid in both the stable structure and the identification of a set of data elements for defining substances in a consistent manner (EMA 2021). Moreover, FDA and the National Center for Advancing Translational Sciences (NCATS) have worked to create the Global Substance Registration System (GSRS) which creates a public access of a database of unique ingredient identifiers and descriptions for active substances, including recombinant proteins, nucleic acids, and small molecule (chemical) drugs, as well as their potential impurities (Peryea et al. 2021).

Global harmonization has always featured prominently in the need for structured data. Indeed, specific International Council for Harmonisation (ICH) activities, and in particular the proposed revision to ICH M4Q, The Common Technical Document For The Registration Of Pharmaceuticals For Human Use: Quality illustrates such an example (ICH 2004). Although ICH M4Q will not necessarily create a format for structured data, the revision will attempt to reorganize the application in a way more suitable to support structured data. Thus, ICH M4Q can potentially better position submissions for assessors using systems that are capable of receiving structured data (e.g., a more helpful version of electronic Common Technical Document (eCTD) Module 2 – Product Quality Summary). Finally, and most notably, a proposal for a future ICH guideline on Structured Product Quality Submissions (SPQS) may provide a mechanism to attempt to standardize a common set of data elements, vocabularies and taxonomies in eCTD Product Quality Module (ICH 2020, 2021). Nevertheless, it is unclear how prescriptive any controlled vocabularies that arise from such initiatives will be for therapeutic proteins.

Critically, each one of these initiatives described serve as a key interrelated piece necessary to modernize regulatory submissions and further improve data standards in the industry and with health authorities. However, a consistent theme of these efforts is the deferral of naming and vocabularies for protein quality attributes to individual entities (e.g., companies, regulatory agencies) and even individual submissions. This limitation, when coupled with the nature of proteins and their typical heterogeneity (and the variety of potential attribute descriptions), threatens to dramatically limit the utility of structured data. As depicted in Fig. 1, the FDA assessment and development of structured data approaches reflect many interconnecting pieces that have been summarized elsewhere (Tran et al. 2024). For example, KASA reflects the internal assessment system which supports the integrated quality assessment, but utilizes and relies on the structured data/information resulting from ICH M4Q, and PQ/CMC. As schematized in the figure, the established controlled vocabularies (and taxonomy)

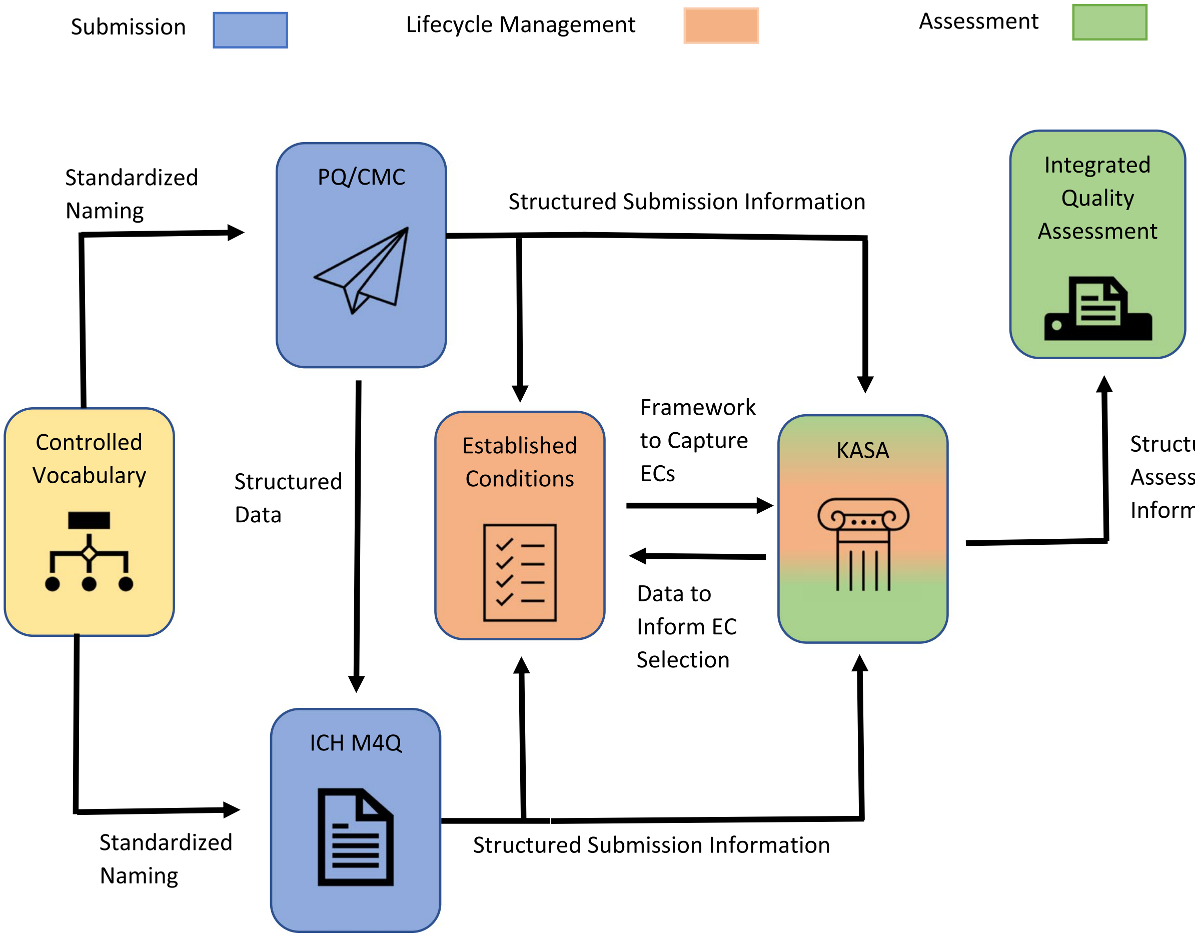


Fig. 1 A pictorial representation of FDA initiatives supporting digital approaches and structured submissions, with a representation of how controlled vocabularies and a corresponding taxonomy play a foundational input role

would be used throughout product lifecycles and consistently throughout these systems. Although this broad concept has been described publicly (Kozlowski 2020), to achieve the intended benefits of structured submissions including more effective use of lifecycle tools such as ICH Q12 (ICH n.d.) and structured data, a structured taxonomical naming of quality attributes is essential and must be a *foundational* building block across all such initiatives for protein products.

Regulatory challenges and opportunities with biologics

There are many factors that drive the inconsistencies observed in protein quality attribute naming and taxonomy. First, biological products have well-documented differences from conventional chemical drugs with respect to their macromolecule size and the number of potential modifications to their structure. Indeed, a key factor that spans all biological products is the heterogeneity of the protein itself, involving a wide variety of different protein attributes (Dougherty et al. 2018). It includes a variability that plays out in batch-to-batch considerations, as well as within batch variability from protein molecule to protein molecule, such as charge and size variants. Secondly, there are dozens of different potential post-translational modifications (PTMs) and with them a multitude of permutations that occur concomitantly, each contributing to the heterogeneity of a therapeutic protein. These modifications may result in variants that retain full activity (i.e., product-related substances) or altered activity (i.e., product-related impurities).

The complexity of protein molecules takes a variety of forms and drives the types of analytical methods and the quantity and breadth of characterization studies that are frequently applied. This is particularly true for biosimilar development as analytics have been long described (and depicted in a pyramid figure) as the foundation on which the entire development resides (Dougherty et al. 2018). This deep analytical characterization demonstrates both the key role that analytics play, and with it, the tremendous amount of information and potential knowledge that could be leveraged and automated in structured data systems. Additionally, exciting advances in biotechnology manufacturing including the development of "platforms" and modular manufacturing allows the opportunity to capture and utilize critical prior knowledge and cumulated experience of a company. Effective utilization and assessment of a platform demands consistent assessment tools, formatting, and critically, structured data to ensure meaningful comparison of a pair of applications and data from a company. Moreover, submission elements unique to biologics such as completed process performance qualification (i.e., PPQ) provided in the original submission are particularly suitable to structured data opportunities and offer an assessor the chance to efficiently compare the process validation data directly with characterization and proposed operation conditions. Finally, proteins may also exhibit indication-specific quality attributes (e.g., importance of a particular effector function for only a subset of indications treated). For this reason, a linkage of a protein product, its quality attributes and its target or pathway are of particular value in structured systems and critical in problem solving such as the identification of an urgent safety signal that may be target specific. A structured system to capture product structural and functional attributes could be combined with clinical and safety information to help predict the impact of structural changes and inform regulatory decisions by applicants and assessors. A summary of the unique nature of a biological product submission and its elements is depicted in Fig. 2.

The complexity of biological product structural characterization, manufacturing processes, data sets and clinical contexts is challenging and at the same time, reflect an opportunity to leverage structured data in predicting risks, capturing and evaluating safety signals early, streamlining assessments and improving regulatory decision making. Unlike small molecule products, biological product modifications, degradants, and variants of the active component do not have such defined naming. Notably, exciting examples of the possibility of real-time algorithmic exchange and processing of pharmaceutical quality data (Anderson et al. 2023) as well as the implementation guides for PQ/CMC do not specify controlled vocabularies for protein products (HL7 International n.d.). Rather, these approaches would currently defer to individual submissions to define them, which tend to be platform-, company- or even application-specific vocabularies. This results in differing naming of the same attribute, the use of similar names to reflect different ensembles of modifications, and collections of names (and their attributes) into categories that are in some cases arbitrary and difficult to predict. This challenge poses a stark limitation to the seamless integration of structured data into computer-based systems. Even if future scenarios could employ AI/ML tools to consider "translation" of quality attribute names in a submission into a consistent structured format, such approaches may be limited or introduce a disconnect between the assessment system and narrative descriptions of quality attributes in the submission. Moreover, an agreed upon format would be beneficial to the regulated industry as well, as it would provide effective ways to organize information that are reproducible and easily understood even for new therapeutic modalities. A hypothetical summary of the potential "unstructured" names in "structured data" for biologics

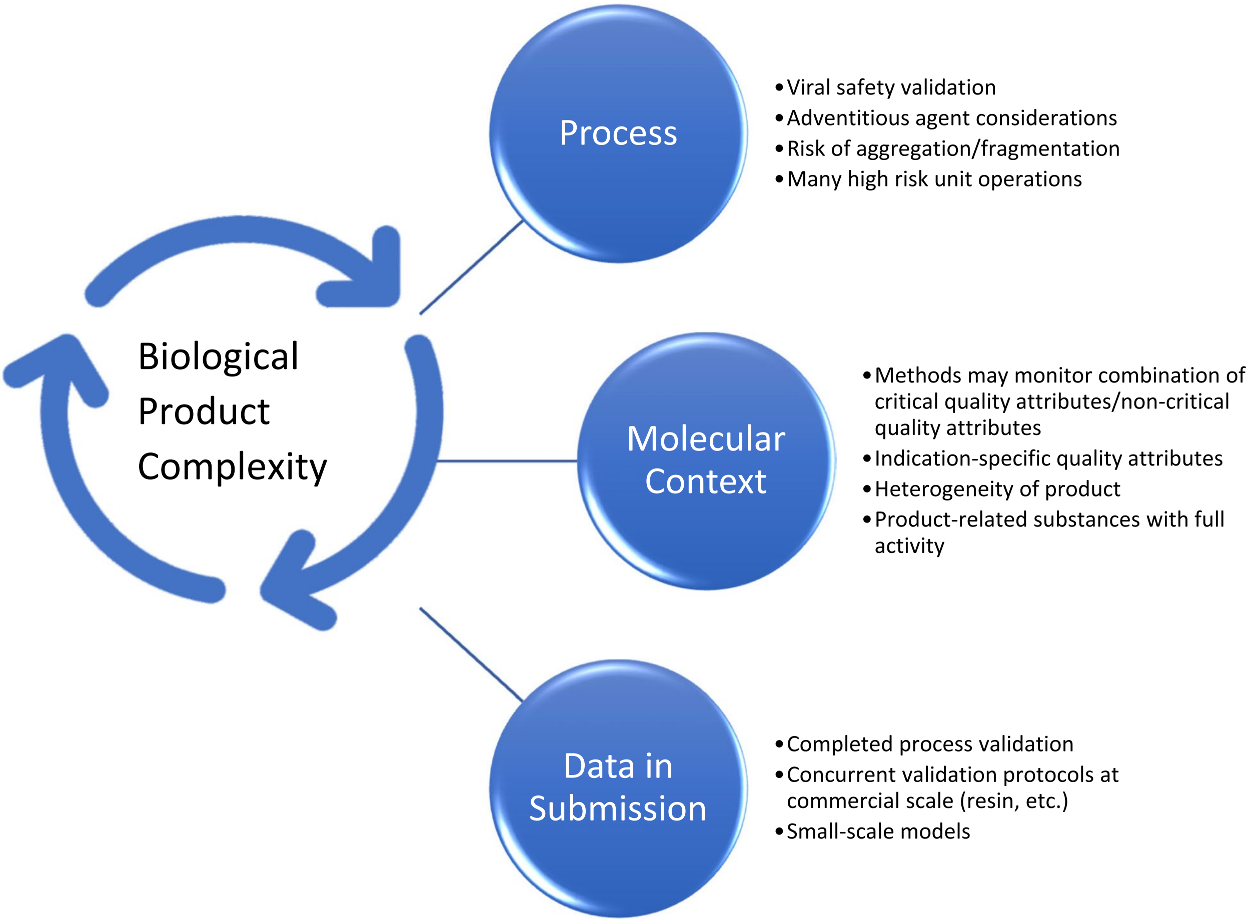

Biological
Product
Complexity
Process
- Viral safety validation
- Adventitious agent considerations
- Risk of aggregation/fragmentation
- Many high risk unit operations
Molecular
Context
- Methods may monitor combination of critical quality attributes/non-critical quality attributes
- Indication-specific quality attributes
- Heterogeneity of product
- Product-related substances with full activity
Data in
Submission
- Completed process validation
- Concurrent validation protocols at commercial scale (resin, etc.)
- Small-scale models

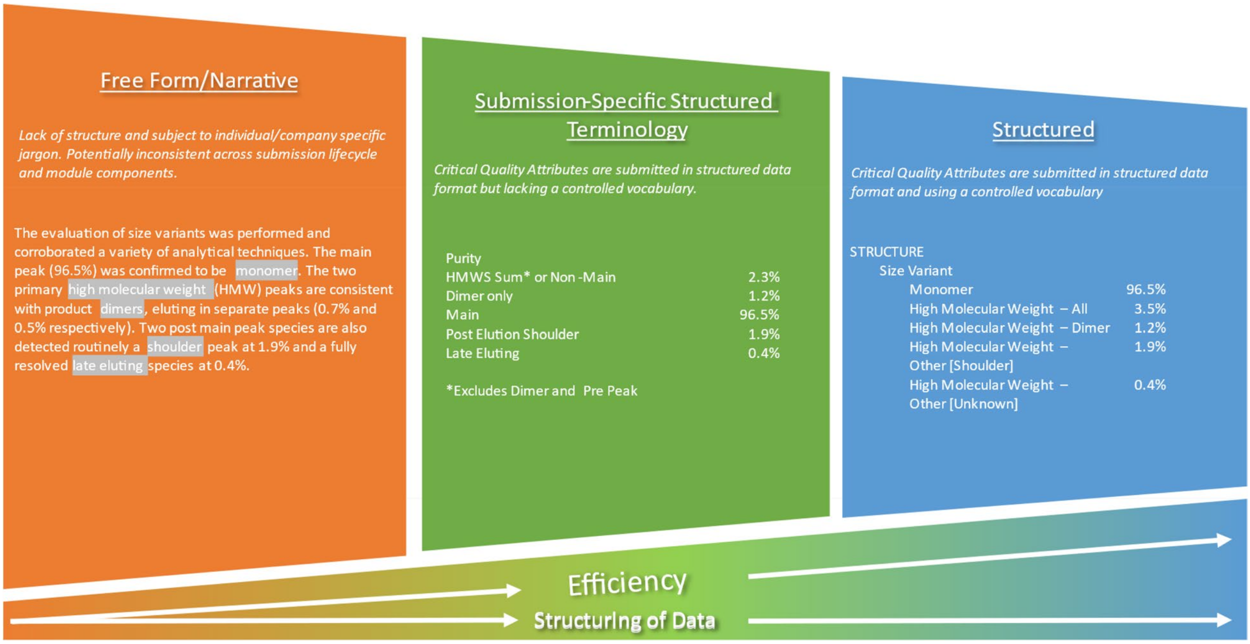
Fig. 2 A schematic depiction of the unique submission elements for a biological product relative to a small molecule

for the example of high molecular weight species is depicted in Fig. 3.

Reflections on conventional approaches

The predominant challenge with the variety of approaches to the vocabulary and taxonomy of quality attributes for protein products is not that any of them are right or wrong. Rather, it is they are merely inconsistent and consequently self-limiting. Indeed, many reasonable and thoughtful scientific approaches have been used to summarize quality attribute data in literature (Alsamil et al. 2020, 2021; Alt et al. 2016; Dash et al. 2021; Vandekerckhove et al. 2018; Tekdemir et al. 2020; Nupur et al. 2022; Zhang et al. 2020; Lee et al. 2018; Saitoh 2018). Nevertheless, there is great deal of inconsistency, and with it, a variability that may manifest itself in a variety of ways. This may include both discrepancies in the approach to placement of a quality attribute within a taxonomic category or inconsistencies with the terminology or approach to defining the terminology of a single attribute itself. Additional sources of variability may also arise from the conflation of the method used to assess the attribute and the actual attribute. These inconsistencies are not surprising given the global nature of drug development, multiplicity in research, the complexity of any single biological product, the diversity of types of biological products under development, and the complex and proprietary nature of manufacturing development.

ICH Q8 (ICH 2009) defines a Critical Quality Attribute as “a physical, chemical, biological, or microbiological property or characteristic that should be within an appropriate limit, range, or distribution to ensure the desired product quality”. It is further clarified that critical quality attributes generally are “associated with drug substance, excipients, intermediate (in-process) and drug product”. Importantly, the terms “property” and “characteristic” are not defined, and understandably may be interpreted slightly differently in their context of use. Moreover, a lack of a generalized, global approach for a naming of attributes has created an inconsistency that results in different approaches that span molecule classes, product portfolios, and even individual applications.


Free Form/Narrative

Lack of structure and subject to individual/company specific jargon. Potentially inconsistent across submission lifecycle and module components.

The evaluation of size variants was performed and corroborated a variety of analytical techniques. The main peak (96.5%) was confirmed to be monomer. The two primary high molecular weight (HMW) peaks are consistent with product dimers, eluting in separate peaks (0.7% and 0.5% respectively). Two post main peak species are also detected routinely a shoulder peak at 1.9% and a fully resolved late eluting species at 0.4%.

Submission-Specific Structured Terminology

Critical Quality Attributes are submitted in structured data format but lacking a controlled vocabulary.

Purity	
HMWS Sum* or Non-Main	2.3%
Dimer only	1.2%
Main	96.5%
Post Elution Shoulder	1.9%
Late Eluting	0.4%

*Excludes Dimer and Pre Peak

Structured

Critical Quality Attributes are submitted in structured data format and using a controlled vocabulary

STRUCTURE	
Size Variant	
Monomer	96.5%
High Molecular Weight – All	3.5%
High Molecular Weight – Dimer	1.2%
High Molecular Weight – Other [Shoulder]	1.9%
High Molecular Weight – Other [Unknown]	0.4%

Fig. 3 A schematic depiction of the efficiency resulting in the transition from free form "pdf" submission to structured data, noting that a truly structured data submission for proteins necessitates a controlled vocabulary

Understandably, given these inconsistencies, a variety of challenges result, including disjointed knowledge management and limited opportunities for informatics. While it is acknowledged that not all proteins share the same quality attributes, we posit that when the quality attributes are the same, they should be named the same and organized in the same way in such applications and even scientific literature. Such a starting point for a quality attribute framework could further be used to refine certain structural terminology, such as that for individual glycans and what structures would be considered to constitute "afucosylation".

The first broad challenge and example of inconsistency results from a conflation of the measurement/method and the quality attribute. A simple such example would be the quality attribute of sterility. Traditionally, it is thought of as the quality attribute, but in some instances, container closure integrity as a surrogate *method* may be reported as a quality attribute. Other quality attributes, such as the measurement of charge variant profile, may broadly reflect a protein property, or it may be used to describe an orthogonal intent to capture a variety of PTMs that are typically associated with the non-main peak in a charge variant profile. Such modifications (e.g., individual deamidations), may be classified and summarized separately from charge or not (Dash et al. 2021), combined with size variants or not (Alsamil et al. 2020; Tekdemir et al. 2020) and separated from oxidation or not (Alt et al. 2016). The conflation may also take other forms, where the technique used to measure an attribute may result in the collection of all attributes into classes for which it is not technically a member, for instance, a non-glycosylated heavy chain may (or may not) be part of glycosylation (Nupur et al. 2022), or purity (Zhang et al. 2020), or further subcategorized as an enzymatic PTM or not a PTM at all (Alsamil et al. 2020). Attribute classification by the overlap of the method and the quality attribute may pose further confusion as technologies evolve and the development of new opportunities in advanced manufacturing. This progress offers the promise of replacing certain contemporary technologies, and with it to challenge collective wisdom on how specific attributes are routinely monitored.

A second challenge with respect to consistent naming how the protein function or biological activity is described at all, in particular, under the term "potency". Potency, as defined in 21 CFR 600.3(s) means "the specific ability or capacity of the product, as indicated by appropriate laboratory tests or by adequately controlled clinical data obtained through the administration of the product in the manner intended, to effect a given result". It is sometimes in submissions used interchangeably with the *potency assay*, which is a specific laboratory method that may be used to provide an assessment of the protein's biological activity for the purpose of batch release and stability monitoring. Notably, therapeutic proteins may have multiple mechanisms of action that are thought to contribute to the activity of product, though not all may be evaluated routinely for a batch release decision for each product lot. Further complicating this approach are biosimilar products, for which such mechanisms are likely to be included as part of the comparative analytical assessment and whose description should match the function and purpose in a predictable way to allow data analysis and informatics necessitated by these quality data rich submissions. Finally, new advanced manufacturing technologies and the possibility of advanced analytics may one day enable broad consideration of alternatives to traditional "cell-based" potency assays and may allow other correlated techniques (such as those used for measuring primary structure modifications or glycosylation) to serve as the predominant control strategy for a protein's biological activity.

A third challenge with consistent nomenclature is that naming approaches sometimes classify the quality attribute based on its relationship to a physiological function. For example, this approach may result in C1q binding being considered bioactivity or not, binding or not, or an immunochemical property or not (Vandekerckhove et al. 2018; Zhang et al. 2020; Lee et al. 2018). This can prove challenging given that certain targets may have physiological functions that result in a wide variety of results in different populations or disease states. Finally, new proteins and scaffolds will certainly challenge the conventional understanding of the breadth of protein function, as well as the traditional approaches used to classify and describe mechanisms of action in their clinical indications.

Lastly, with respect to consistency of placement in a hierarchy (i.e., a taxonomy) some traditional approaches may collect the quality attributes into groups that are not scientifically driven but reflect a regulatory risk process platform and company's approach at a point in time, for example, "obligatory critical quality attributes" (Saitoh 2018). Such groupings are understandable and may generically describe a wide range of attributes, and could be applicable specifically to drug substance and/or drug product (Vandekerckhove et al. 2018; Lee et al. 2018). Notably, as a recent survey demonstrated, such approaches vary dramatically from company to company (Demmon et al. 2020).

Key aspects of the proposed approach

As presented and summarized in Fig. 4, we have designed a taxonomy and controlled vocabulary framework that can be used across therapeutic proteins. This framework can be used for all therapeutic proteins independent of product indication, drug product presentation, or phase in drug development. Notably, this framework as presented in the figure does not include all subcategories at the third sub-category level (i.e., cardinality level 0.01) or fourth level (i.e., cardinality level 0.001). This is both a reflection of a logistical limitation of presentation in a figure and realistic acknowledgement that such detail is not entirely knowable for such a large portfolio and evolving group of biological products. Nevertheless, it creates an expandable approach that can be adapted for new products as they are developed and allow for predictable locations and categorization of quality attributes. In the following section, we define and describe the key principles that inform our proposed approach:

- Principle 1: A Top-Down View of the Product and Protein Structure
- Principle 2: A Decoupling of a Quality Attribute and the Test Used to Evaluate It
- Principle 3: Adaptability for New Emerging Products and New Technologies
- Principle 4: Classification of Biologic Activity into Clear and Indication Agnostic Outcome
- Principle 5: Implementable Across All Dossier Elements
- Principle 6: Allows Focus to Capture Necessary Quality Attributes Only

Our controlled vocabulary begins with six main categories at the highest level (cardinality 1.0): Active Ingredient, Structure, Function, Process-Related Impurities, Material Properties, and Formulation. These highest-level categories serve to establish our first two key principles: taking a top-down view of the protein itself that separates the whole protein from the other parts of the product. It also ensures that the quality attribute and the method used to measure it do not become conflated. Further, this approach distinguishes whole molecular properties which we refer to as active ingredient properties (such as charge and mass) and structural properties (e.g., primary structure). Specifically, this aids in deconvoluting measurements that are closely related and overlapping (e.g., charge, which is related to but distinct from the primary structure modifications that it may detect, such as deamidation).

The structure category at level 1.0 contains all modifications to the structure of the active ingredient, including: higher order structure, primary structure, size variants, and linked non-protein polymer. Linked non-protein polymer includes all covalent modification to the amino acid polymer, including conjugation and glycosylation. This allows a predictable approach for new molecular constructs, as well as for organizing familiar molecules that have been manufactured in an unconventional way. For example, cell-free protein synthesis approaches may allow opportunities to target creation of difficult to express proteins and offers chances to

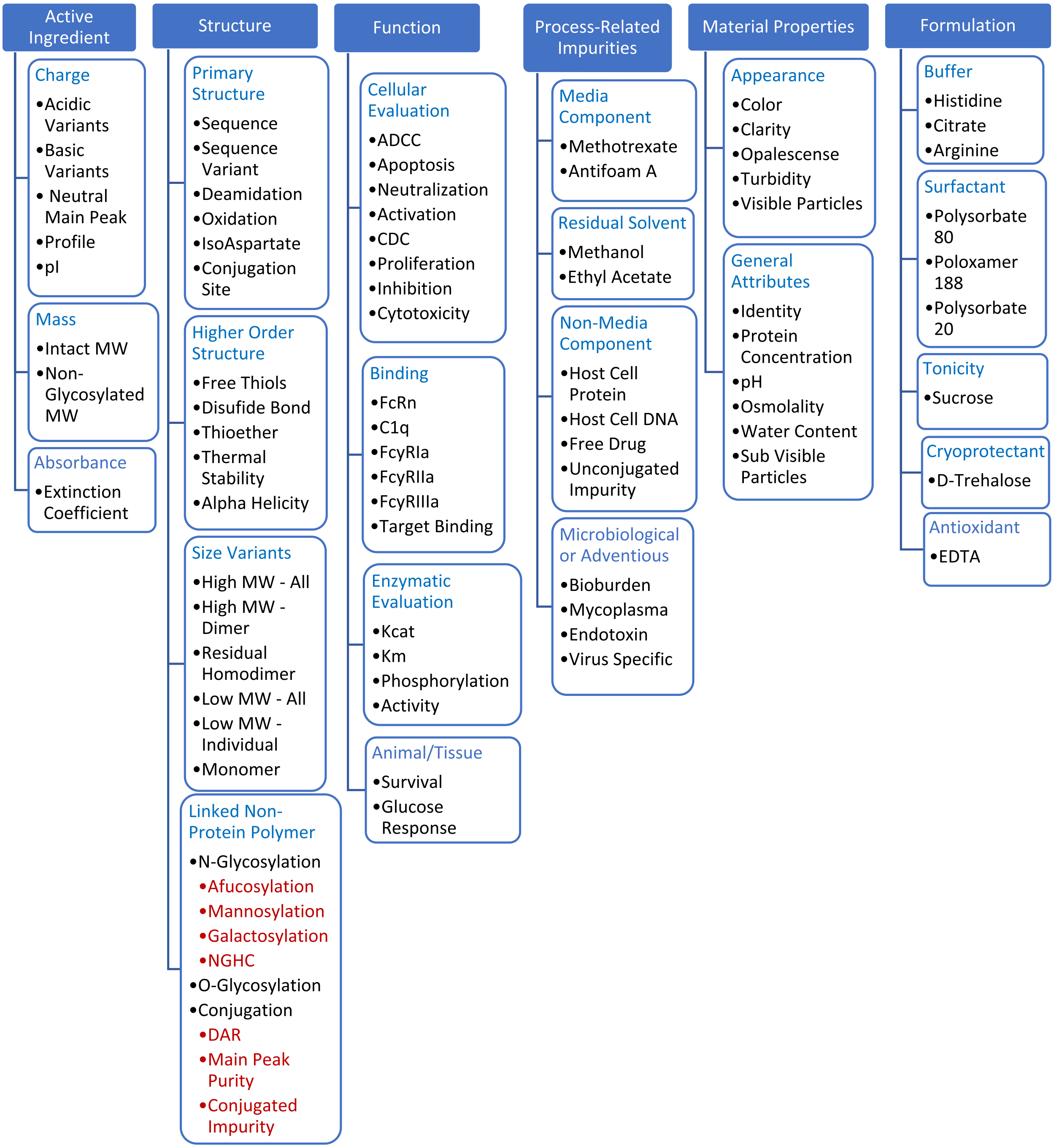


Fig. 4 Design of proposed taxonomy and controlled vocabulary framework. The main categories at cardinality 1.0 are shown in blue boxes. Notably, this framework as presented in the figure does not include all possible examples beyond the second cardinality level (0.1), which is presented in blue font. The third level cardinality is presented in black font (0.01). The fourth level cardinality (0.001) applies only to the Linked non-protein polymer cardinality of 0.01 and is depicted in red font

incorporate non-natural amino acids, and targeted glycosylation approaches, which challenge conventional terminology and the intuitive meaning of current vernacular such as "post-translational modifications" (Chiba et al. 2021). Furthermore, we feel this categorization also supports future looking program development for new molecules and manufacturing methods (Shatz et al. 2016). Such types of new approaches will need predictable locations for quality attributes and their variants that arise in ways that differ from our typical classification and familiarity.

A key feature that differentiates biological products generally from small molecule drugs is describing its biological function(s). As noted in Reflections on conventional approaches section, this description is perhaps the greatest variability and challenge, as a consequence, in the implementation of structured data. For this reason, our naming taxonomy attempts to decouple the readout, the assay parameters, and the target. Rather, we propose to include specific subclasses that offer cellular evaluation, enzymatic evaluation, or a specific binding event that will allow predictable categories. Where needed, "animal/tissue" and more specific evaluation (e.g., survival) could be added to define cardinality levels 0.1 and 0.01 for this subclass for certain products which still utilize those types of quality attributes from such living systems (i.e., in vivo measurement). This general approach may prove particularly amenable for future applications where advanced manufacturing technologies and the maturity of advanced analytics may one day allow for broad consideration of alternatives to contemporary assays for routine control. Moreover, future constructs such as multi-specific antibodies offer such potential opportunities to better tune activity and with it, introduce a wider variety of potential features and quality attributes that need to be described consistently.

We acknowledge that all models have limitations. Indeed, our framework itself has certain obvious considerations that need to be acknowledged. First, like any framework, locations of certain attributes may not be intuitive, especially relative to historical practice. For example, heavy chain antibody species lacking glycosylation (also referred to as NGHC) is located in the N-glycosylation group at the cardinality 0.01 in the subcategory (0.001) of Linked non-protein polymer in lieu of with antibody size variants, with which it is typically measured analytically. Additionally, the location within the taxonomy will require an emphasis on its context rather than not its name or origin. For example, an assessor would need to evaluate whether the impurity in an antibody-drug conjugate is conjugated the active ingredient (Structure / Linked Non-Protein Polymer / Conjugated Impurity) or not (Process-Related Impurity / Non-Media Component / Unconjugated Impurity). Nevertheless, with predictability and clarity, such shifts as in these subcategorizations of attributes would be easily manageable.

The successful development of a naming taxonomy (including controlled vocabulary) and the consistency and predictability of the approach we envision supports two final additional key principles: flexibility for applicants to identify necessary quality attributes, and a vocabulary that can be used throughout an entire application. This approach generates consistency in sections related to characterization, specification, stability, and analytical similarity (where applicable). We acknowledge that some quality attributes, in particular, those related to critical patient safety (e.g., presence of adventitious agents) and for which a quantitative specification may never be established (as their presence would instead result in a complete response letter), may still technically be considered quality attributes. The ultimate utility of the identification and submission of quality attributes is in a dossier along with description in sections such as specifications, stability, and product characterization. For this reason, we feel it is preferable for a taxonomy to place such attributes within predictable categories based on scientific principles in lieu of generalizing them into such groups as "obligatory" or "mandatory". This approach allows companies to include them (or not) based on the needs and context of their product and its control strategy.

Conclusions

Herein we have provided a model for organizing quality attributes of therapeutic proteins in a harmonized and structured way. The intent is not that such a naming taxonomy and vocabulary will provide a prescriptive approach, but rather that it may spark a coordinated effort to synergize with other structured data efforts already underway. This effort would be rooted in both a desire to find a collective approach, but also an acknowledgement that a unified approach is a must have to ensure the success of structured data for biological products in an era of digitalization. Moreover, while individual category names may be further refined, our hope and vision is that key scientific and regulatory principles of a predictable approach will position regulatory dossiers to interface seamlessly with structured data submissions and the assessment systems that house them.

Abbreviations

AI	Artificial Intelligence
ANDA	Abbreviated New Drug Application
BLA	Biologics License Application
CDER	Center for Drug Evaluation and Research
CPP	Critical Process Parameters
CQA	Critical Quality Attributes
CTD	Common Technical Document
DAR	Drug-Antibody Ratio
DLD	Drug Load Distribution
EC	Established Conditions
EMA	European Medicines Agency
FDA	Food and Drug Administration
FHIR	Fast Healthcare Interoperability Resources
GSRS	Global Substance Registration System
HL7	Health Level 7
ICH	International Conference for Harmonisation
IDMP	Identification of Medicinal Products
ISO	International Organization for Standardization
KASA	Knowledge Aided Assessment and Structured Application
ML	Machine Learning
MW	Molecular Weight

NGHC	Non-Glycosylated Heavy Chain
NDA	New Drug Application
PQ/CMC	Pharmaceutical Quality/Chemistry Manufacturing and Controls
PPQ	Process Performance Qualification
PTM	Post-Translational Modification
SPQS	Structured Product Quality Submission

Acknowledgements
The authors would like to thank Ashutosh Rao, Cristina Ausin, Nina Brahme, Stacey Ricci, Leslie Rivera-Rosado and Joanna Qing Zhou for their thoughtful suggestions throughout the development of the manuscript.

Authors' contributions
All authors contributed to the developing, writing, and the editing of the manuscript submitted. All authors read and approved the final manuscript.

Funding
The authors did not receive any specific grant from funding agencies in the public, commercial or not-for-profit sectors.

Availability of data and materials
Not applicable.

Declarations

Competing interests
The authors declare that they have no competing interests.

Received: 29 April 2024 Accepted: 21 June 2024
Published online: 01 August 2024

References

Ahluwalia K, Abernathy MJ, Beierle J, Cauchon NS, Cronin D, Gaiki S, Lennard A, Mady P, McGorry M, Sugrue-Richards K, Xue G (2022) The future of CMC regulatory submissions: streamlining activities using structured content and data management. J Pharm Sci 111(5):1232–1244. https://doi.org/10.1016/j.xphs.2021.09.046

Algorri M, Cauchon NS, Abernathy MJ (2020) Transitioning chemistry, manufacturing, and controls content. J Pharm Sci 109(4):1427–1438. https://doi.org/10.1016/j.xphs.2020.01.020

Algorri M, Abernathy MJ, Cauchon NS, Christian TR, Lamm CF, Moore CM (2022) Re-envisioning pharmaceutical manufacturing: increasing agility for global patient access. J Pharm Sci 111(3):593–607. https://doi.org/10.1016/j.xphs.2021.08.032

Allotrope Foundation A (n.d.) https://www.allotrope.org. Accessed 15 Apr 2024

Alsamil AM, Giezen TJ, Egberts TC, Leufkens HG, Vulto AG, van der Plas MR, Gardarsdottir H (2020) Reporting of quality attributes in scientific publications presenting biosimilarity assessments of (intended) biosimilars: a systematic literature. Eur J Pharm Sci 154:105501. https://doi.org/10.1016/j.ejps.2020.105501

Alsamil AM, Giezen TJ, Egberts TC, Leufkens HG, Gardarsdottir H (2021) Type and extent of information on (potentially critical) quality attributes described in European public assessment reports for adalimumab biosimilars. Pharmaceuticals 14(3):189. https://doi.org/10.3390/ph14030189

Alt N, Zhang T, Motchnik P, Taticek R, Quarmby V, Scholthauer T, Beck H, Emrich T, Harris RJ (2016) Determination of critical quality attributes for monoclonal antibodies. Biologicals 44(5):291–305. https://doi.org/10.1016/j.biologicals.2016.06.005

Anderson C, Algorri M, Abernathy M (2023) Real-time algorithmic exchange and processing of pharmaceutical quality. Int J Pharm 645:123342. https://doi.org/10.1016/j.ijpharm.2023.123342

Arden NS, Fisher AC, Tyner K, Yu LX, Lee SL, Kopcha M (2021) Industry 4.0 for pharmaceutical manufacturing: preparing for the smart factories of the future. Int J Pharm 602:120554. https://doi.org/10.1016/j.ijpharm.2021.120554

Beierle J, Algorri M, Cortes M, Cauchon NS, Lennard A, Kirwan JP, Abernathy MJ (2023) Structured content and data management enhancing acceleration in drug development through efficiency in data exchange. AAPS Open. https://doi.org/10.1186/s41120-023-00077-6

Brennan Z (2018) End of the eCTD? FDA pushes for new KASA system to improve assessments. Retrieved from https://www.raps.org/news-and-articles/news-articles/2018/9/end-of-the-ectd-fda-pushes-for-new-kasa-system-to. Accessed 15 Apr 2024

Cauchon NS, Oghamian S, Hassanpour S, Abernathy M (2019) Innovation in chemistry, manufacturing, and controls-a regulatory perspective from industry. J Pharm Sci 108(7):2207–2237. https://doi.org/10.1016/j.xphs.2019.02.007

Chiba CH, Knirsch MC, Azzoni AR, Moreira AR, Stephano MA (2021) Cell-free protein synthesis: advances on production process for biopharmaceuticals and immunobiological products. Biotechniques 70(2):126–133. https://doi.org/10.2144/btn-2020-0155

Cox B (2021) Newly aligned teams sped US FDA's drug quality reviews over pandemic hurdles. Retrieved from Pink Sheet: https://pink.pharmaintelligence.informa.com/PS143801/Newly-Aligned-Teams-Sped-US-FDAs-Drug-Quality-Reviews-Over-Pandemic-Hurdles. Accessed 15 Apr 2024

Dash R, Singh SK, Chirmule N, Rathore AS (2021) Assessment of functional characterization and comparability of biotherapeutics: a review. APPS J 24(1):15. https://doi.org/10.1208/s12248-021-00671-0

Demmon S, Bhargava S, Ciolek D, Halley J, Jaya N, Joubert MK, Koepf E, Smith P, Trexler-Schmidt M, Tsai P (2020) A cross -industry forum on benchmarking critical quality identification and linkage to process characterization studies. Biologicals 9–20. https://doi.org/10.1016/j.biologicals

Dougherty M, Zineh I, Christl L (2018) Perspectives on the current state of the biosimilar regulatory pathway in the United States. Clin Pharmacol Ther. https://doi.org/10.1002/cpt.909

EMA (2021) Products management services - implementation of International Organization for Standardization (ISO) standards for the Identification of Medicinal Products (IDMP) in Europe. Retrieved from https://www.ema.europa.eu/en/documents/regulatory-procedural/guideline/products-management-services-implemenation-international-organization-standardization-iso-standards_en.pdf. Accessed 15 Apr 2024

FDA (2019) Pharmaceutical Quality/Chemistry Manufacturing and Controls (PQ/CMC) data elements and terminologies. Retrieved from https://www.regulations.gov/document/FDA-2017-N-2166-0001. Accessed 15 Apr 2024

FDA (2023) PQ/CMC and IDMP. Retrieved from https://www.fda.gov/industry/pharmaceutical-quality-chemistry-manufacturing-controls-pqcmc/pqcmc-and-idmp. Accessed 15 Apr 2024

HL7 International (n.d.) PQ/CMC FHIR implementation guide version 0.1.19. https://build.fhir.org/ig/HL7/FHIR-us-pq-cmc-fda/. Accessed 15 Apr 2024

ICH (2004) M4Q(R1) CTD on quality. Retrieved from https://database.ich.org/sites/default/files/M4Q_R1_Guideline.pdf. Accessed 15 Apr 2024

ICH (2009) Q8(R2) pharmaceutical development. https://database.ich.org/sites/default/files/Q8%28R2%29%20Guideline.pdf. Accessed 15 Apr 2024

ICH (2020) ICH press release 2020. Retrieved from https://admin.ich.org/sites/default/files/2020-06/ICH40MayTC_PressRelease_2020_0603_FINAL_0.pdf Accessed 15 Apr 2024

ICH (2021) 2020 annual report. Retrieved from https://admin.ich.org/sites/default/files/inline-files/ICH_AnnualReport_2020_2021_0602.pdf. Accessed 15 Apr 2024

ICH (n.d.) Q12 Technical and regulatory considerations for pharmaceutical product lifecycle management. Retrieved from: https://database.ich.org/sites/default/files/Q12_Guideline_Step4_2019_1119.pdf. Accessed 15 Apr 2024

Kozlowski S (2020) A shared framework for protein analytics; bioassays enhancing drug development. CASSS Bioassay Presentation. Retrieved from https://www.casss.org/papers-and-presentations/resource/a-shared-framework-for-protein-analytics-bioassays-enhancing-drug-development. Accessed 15 Apr 2024

Lee J, Kang HA, Soo Bae J, Kim KD, Lee KH, Lim KJ, Choo MJ, Chang SJ (2018) Evaluation of analytical similarity between trastuzumab biosimilar CT-P6 and reference product using statistical analyses. MAbs 10(4):547–571. https://doi.org/10.1080/19420862.2018.1440170

Macdonald JC, Isom DC, Evans DD, Page KJ (2021) Digital innovation in medicinal product regulatory submission, review, and approvals to create a dynamic regulatory ecosystem—are we ready for a revolution? Front Med (Lausanne) 8:660808. https://doi.org/10.3389/fmed.2021.660808

Nupur N, Joshi S, Guilliarme D, Rathore AS (2022) Analytical similarity assessment of biosimilars: global regulatory landscape, recent studies and major advancements in orthogonal platforms. Front Bioeng Biotechnol 10:832059. https://doi.org/10.3389/fbioe.2022.832059

Peryea T, Southall N, Miller M, Katzel D, Anderson N, Neyra J, Stemann S, Nguyen DT, Amugoda D, Newatia A, Ghazzaoui R, Johanson E, Diederik H, Callahan L, Switzer F (2021) Global Substance Registration System: consistent scientific descriptions for substances related to health. Nucleic Acids Res 49(D1):D1179–D1185. https://doi.org/10.1093/nar/gkaa962

Pistoia Alliance (n.d.) https://www.pistoiaalliance.org. Accessed 15 Apr 2024

Robertson AS, Malone H, Bisordi F, Fitton H, Garner C, Holdsworth S, Honig P, Hukkelhoven M, Kowalski R, Milligan S, O'Dowd L, Roberts K, Rohrer M, Stewart J, Taisey M, Thakkar R, Van Baelen K, Wegner M (2019) Cloud-based data systems in drug regulation: an industry perspective. Nat Rev Drug Discov 19(6):365–366. https://doi.org/10.1038/d41573-019-00193-7

Rosencrance S, Raw A, Smith D, Slack MA (2019) FDA's new initiative. KASA. Retrieved from https://pqri.org/wp-content/uploads/2019/04/PQRI_KASA-Presentation_V4.pdf. Accessed 14 Apr 2024

Saitoh S (2018) The identification of critical quality attributes for the development of antibody drugs. Pharm Soc Jpn 38(12):1475–1481. https://doi.org/10.1248/yakushi.18-00020-1

Schmuff NR (2019) 2nd public meeting on PDUFA VI electronic submissions and data standards: structured PQ/CMC. Retrieved from https://www.fda.gov/media/124614/download. Accessed 15 Apr 2025

Shatz W, Ng D, Dutina G, Wong AW, Dunshee DR, Sonoda J, Shen A, Scheer JM (2016) An efficient route to bispecific antibody production using single-reactor mammalian co-culture. MAbs 1487–1497. https://doi.org/10.1080/19420862.2016.1234569

Tekdemir ZB, Seckin AI, Kacar T, Yilmaz E, Bekiroglu S (2020) Evaluation of structural, biological, and functional similarity. Pharm Res 37(11):215. https://doi.org/10.1007/s11095-020-02932-7

Tran R, Fraser G, Fisher AC, Lee SL, Boam A, Tsinotides S, Maguire J, Yu LX, Rosencrance S, Kozlowski S, Henry D (2024) A network of regulatory innovations to improve FDA quality assessments of human drug applications. Int J Pharm X 7. https://doi.org/10.1016/j.ijpx.2024.100239

Vandekerckhove K, Seidl A, Gutka H, Kumar M, Gratzl, Keire D, Coffey T, Kuehne H (2018) Rational selection, criticality assessment, and tiering of quality attributes and test methods for analytical similarity evaluation of biosimilars. AAPS J 20(4):68. https://doi.org/10.1208/s12248-018-0230-9

Zhang E, Xie L, Qin P, Lu L, Yanpeng X, Gao W, Wang L, Xie MH, Jiang W, Liu S (2020) Quality by design-based assessment for analytical similarity of adalimumab biosimilar HLX03 to Humira. AAPS J 22(3):69. https://doi.org/10.1208/s12248-020-00454-z

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.