Machine Learning and the Trial Master File

Improving Productivity and TMF Inspection-Readiness

Machine learning can quickly classify and index documents for filing in an eTMF. It also could potentially eliminate the need for time-consuming file exchanges that transfer documents between eTMFs. The result: simplifying processes and reducing costs.

Keefer Consulting, Inc.

Contents

Keefer Consulting, Inc.
PO Box 54486
Philadelphia, PA 19148
(215) 462-1601
www.keeferconsulting.com

EXECUTIVE SUMMARY

Keeping a trial master file (TMF) ready for regulatory inspections is critical to avoiding negative TMF findings. Maintaining inspection-readiness requires the timely filing of all trial records, yet current methods for filing trial documents are inefficient and error-prone. Getting documents into an electronic TMF (eTMF) system is largely manual, often involving several hand-offs.

Uploading documents from external sources including investigator sites, clinical research organizations (CROs), other service providers, and regulatory authorities can be resource-intensive and take many weeks to complete. Preparing for an inspection can involve project resources to locate documents missing from the TMF.

Companies must adopt better processes so that study teams are able to focus on managing successful trials. One problem has been simply identifying what a document is. Is it a standard form issued by the FDA or other agency? To what study does it pertain? From which country was it issued? Does the document pertain to a particular site?

Traditional methods require someone to examine each document while answering questions like these. The reviewer must identify and understand metadata terms embedded in document text. Metadata describes specific document attributes such as the study to which the document pertains. eTMF systems require metadata in their databases to help identify and locate documents. Often, someone must enter metadata manually into the eTMF for each document filed.

Machine learning (ML) is one solution emerging to streamline TMF management. As a form of artificial intelligence (AI), ML can reduce manual intervention and associated costs. It can accelerate the classification and indexing of documents for filing them in an eTMF. It could potentially eliminate time-consuming exchanges of trial documents between eTMF systems.

Adoption of the electronic Trial Master File Exchange Mechanism Standard (eTMF-EMS) by eTMF vendors could further extend these benefits. Regardless of the pace of eTMF-EMS adoption, the benefits of leveraging current ML technology compel trial sponsors and CROs to rethink how they process and store TMF documents.

GETTING DOCUMENTS INTO AN ELECTRONIC TRIAL MASTER FILE

Creating a document in an electronic trial master file (eTMF) has traditionally required one or more individuals to first examine the source document or a copy of that document. This involves recognizing the kind of document, or *artifact*, it is and classifying it with similar documents. It is also necessary to create an index to associate the artifact to the appropriate study and indicate whether or not it pertains to a particular country and/or site.

Documents originate frequently from within organizations or from outside sources such as investigator sites, clinical research organizations (CROs), other service providers, and regulatory authorities. Adding this content to an electronic trial master file and keeping the TMF inspection-ready requires ongoing vigilance.

Paper documents are scanned and converted to a digital format, primarily the *PDF* format. A PDF file presents images of text and graphics as they appear on the pages of a paper document. PDFs may be loaded directly into an eTMF or any system that manages documents. The system stores scanned documents as individual files or as files in electronic folders.

Unscanned documents may be in different formats depending on the applications in which they are created. For example, a document may be in a *.docx* file, which is the native format used by Microsoft Word. Alternatively, a native file may be converted to a PDF. The creator of the document may place the file in an electronic folder with other documents, each of which may have a different format. Sometimes document files and/or folders are transferred to an eTMF from another system through a *file exchange*.

Document text often contains terms that may serve as *metadata*, or data that provides information about other data to facilitate its use by humans and computers. Information provided by metadata includes data attribute names, their meaning, and formatting information. eTMF systems require metadata in their databases to help identify and locate particular documents.

The person responsible for indexing the document must be able to recognize the metadata necessary to construct the index. He or she may infer a metadata term if it does not occur explicitly in the text. This task requires experience and awareness of the context in which the document exists. For example, one could infer the identity of a study when the CRO

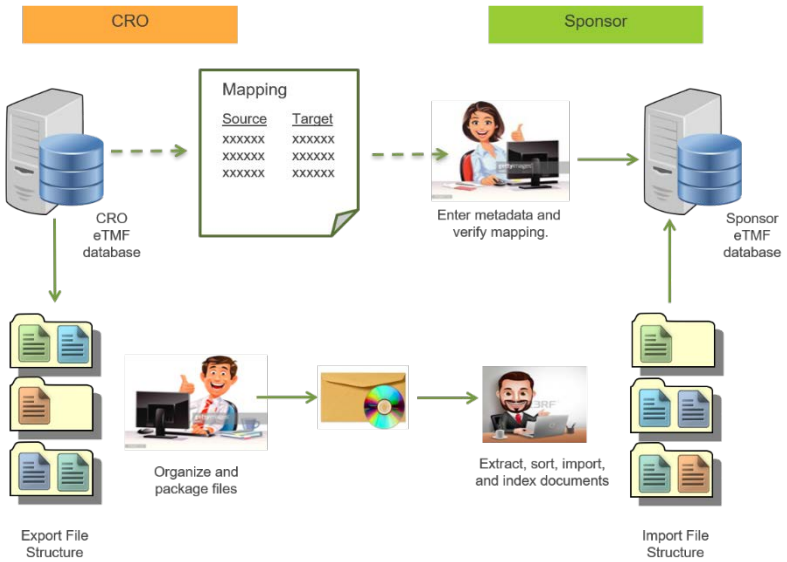

Figure 1 - A typical file exchange using current methods

providing a document is managing just one study for the sponsor.

Metadata creation in an eTMF is often manual but is sometimes automated with a computer program that is customized for a particular file exchange. When transferred from another system through a file exchange, it is necessary to map the metadata terms used by the sending party to the terms used by the receiving party. Mapping metadata is resource-intensive and can take many weeks to complete. Figure 1 shows a typical file exchange process using current methods.

SOLUTION CRITERIA

The envisioned solution to these problems would improve productivity through several capabilities:

- Accelerate the classification and indexing of documents for filing in an eTMF
- Eliminate document file exchanges and the metadata mapping they require
- Reduce costs of manual intervention

MACHINE LEARNING AND THE TMF

Companies are beginning to use *machine learning* (ML) to aid the classification and indexing of documents filed in an eTMF. Automated support for these functions greatly reduces the time and manual effort that has been traditionally necessary.

The technology can eliminate the transfer of documents from non-eTMF systems by supporting the filing of copies in both the document source system and the eTMF. Eliminating file exchanges reduces time and manual effort by eliminating the need to map metadata from one system to another.

ML is a type of artificial intelligence (AI) which aims to simulate human cognition and behavior. Machine learning employs standard computer algorithms that learn and improve from experience. Training an ML algorithm to recognize document features and the patterns in which they occur results in a classification model. The model maps various features and patterns to specific types of documents in which they appear. Once the algorithm recognizes the appropriate document type, it can use metadata extracted from the document by optical character recognition (OCR) to create a document index.

ML differs from *expert systems*, another type of AI. Expert systems are programmed to respond in predetermined ways to specific if/then conditions. Expert systems work well when the number of conditions an application will need to handle is small and known in advance. ML works well when it is impractical to anticipate every possible condition. ML is appropriate for applications that make repeated decisions about frequently occurring use cases.

For example, regulatory authorities receive submissions of standard forms from sponsors and CROs (e.g., FDA Form 1572). While agencies receive certain types of forms repeatedly, their content varies. A response may be handwritten

or printed. Font types and sizes may vary. The scanning quality may affect the appearance of a document. With ML, the algorithm may learn that a certain field appears consistently in the same place on a specific page of a standard form. The ability of an algorithm to recognize specific types of forms improves with the number of examples it processes.

The solution is not limited to the preceding example, as many documents in an eTMF are not standard forms. An algorithm may learn to recognize a certain phrase that indicates the document is a letter from a regulatory authority. ML applications have improved performance in processing many document formats, including email. While some non-standard documents (e.g., meeting minutes) are generally more difficult for ML to process, the technology is continually advancing.

Why now?

Trends supporting the growth of ML include:

Availability of powerful hardware and software, especially in the cloud: Cloud computing offers economies of scale and easier integration of global enterprise computing resources. Companies can avoid the costs of on-premise software (e.g., data centers, IT staff, downtime).

Standardization of algorithms: Mobile computing has sparked new growth in creativity. Algorithms now exist for many standard use cases, including the detection of patterns in images, classification of images and text, and text summarization.

Improvement in data quality management: ML produces better outputs when data inputs are consistently accurate and complete. Statistical techniques to ensure data cleanliness have

Possible Machine Learning Functions

Triage a document: Identify the file format of a document and determine how to process the document before classifying and extracting data from it.

Digitize a scanned document: Translate a scanned document into logical records that a computer program can read (both text and images). Assess whether the document is fit for OCR processing, or if it must first be rescanned. Execute OCR and produce the digital output records.

Standardize the data structure: Convert a document into a standard file format such as XML that allows the transmission of both content and metadata. The document input for the conversion may be in a native format or consist of records produced by digitizing a scanned document.

Classify a document: Determine the document type according to a defined classification scheme, such as the TMF Reference Model [2]. The function may include a feedback loop, allowing the user to review an automatically determined classification. The user teaches the algorithm how the algorithm should classify similar documents by either confirming or correcting the determined classification.

Extract data from document fields: Data in document fields may become parts of a key used to index the document in a database (e.g., Trial, Country, and Site). Extracted attributes may also include those that provide information of value for other operational or analytical uses.

Forward output to a database: A user interface provides options for approving the document and its metadata for addition to the database of an eTMF or other document management system. Statistical methods may calculate confidence values to assist with quality control. Extracted data attributes could also be directed to an application, such as a Clinical Trial Management System (CTMS), either automatically or with user review and approval.

helped improve data quality. Companies are growing increasingly skilled in using these tools.

Availability of examples for training ML: *Data coverage* means having enough scenarios to build a robust ML model. It is impractical to envision every possible type of regulatory form, style of handwriting, or error. Having many examples serves as a proxy for data coverage.

Regulatory authorities have required companies to produce and maintain greater amounts of documentation for clinical trials. Many examples of various types of documents are now available for training ML algorithms to classify document types, identify extracted metadata, and index content.

Training the Algorithm

A document classification model identifies types of documents based on common patterns. Patterns consist of features that an algorithm can observe in a document. Features may include the page format (e.g., letter, A4, etc.) or the location of a signature on a page. The model maps features like these to specific document types, or *labels*. Examples of eTMF document labels include Form 1572, email, meeting minutes, etc.

A model is developed by training an algorithm to recognize patterns, assign labels to classify documents, and use metadata extracted from documents by optical character recognition (OCR) to construct indexes for filing documents in an eTMF. The training data set consists of many examples of different types of documents and metadata. Possible sources of training data include existing systems and vendor libraries.

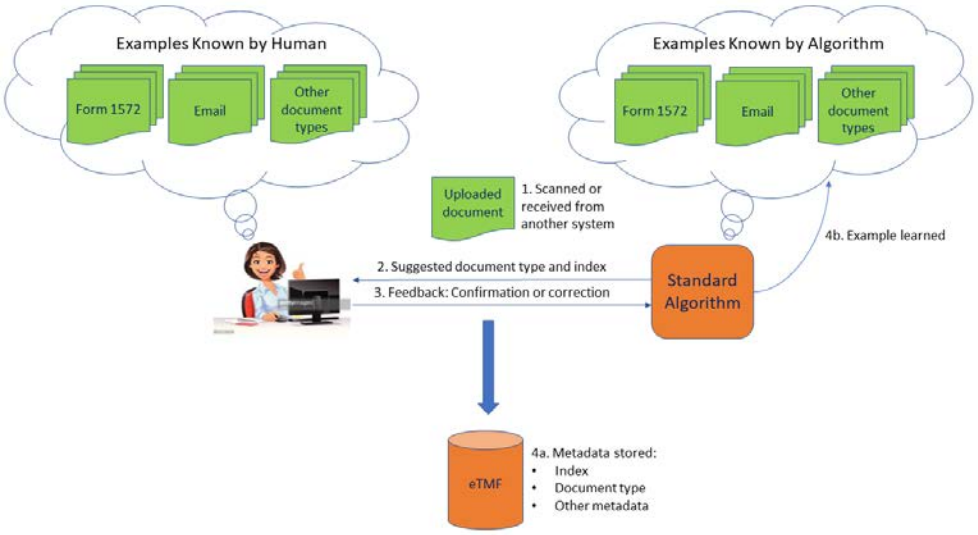

Figure 2 - Training with Human Feedback

Validation

An ML application may include a feedback loop through which a human can confirm or correct a classification label or index suggested by the algorithm (Figure 2). The algorithm uses feedback to improve its classification and indexing performance. ML can learn to recognize some types of documents more easily than others. A trained algorithm can recognize at least some document types reliably at high levels of precision. These documents may bypass the loop once the algorithm has met an acceptable target precision threshold.

Developers of models initially use feedback loops while training an algorithm. Sponsors may use a feedback loop when striving for 100% accuracy to comply with 21 CFR 11. A risk-based quality management approach may allow certain document types to bypass feedback. An algorithm could automatically process types of documents for which precision is validated

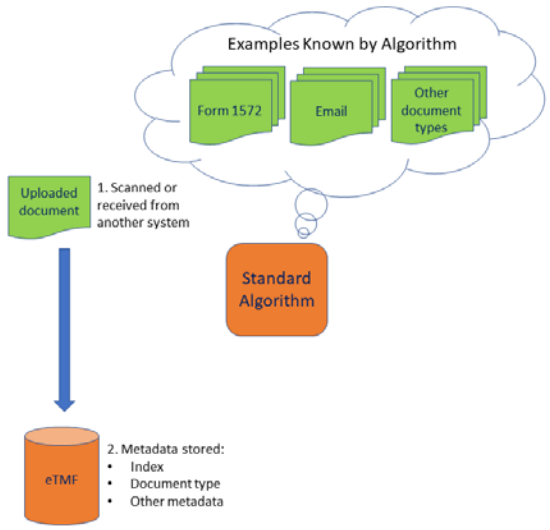

Figure 3 - Operating without Human Feedback

against an acceptable threshold (Figure 3).

The data on which an ML model has been trained provides the standard for testing the algorithm. Variation in different features changes as new documents are processed over time. There is an occurrence of low-quality data such as incorrect classification labels and metadata in many environments. Periodic retesting with an expanded library of quality training examples becomes necessary to maintain precision in classification and indexing. Quality improvement requires manual validation and correction of data before adding it to the training data set. [1]

The electronic Trial Master File Exchange Mechanism Standard (eTMF-EMS)

The TMF Reference Model names and defines documents commonly used in managing clinical trials. The electronic Trial Master File Exchange Mechanism Standard (eTMF-EMS) extends this model. [2]

eTMF-EMS allows software vendors to map proprietary data structures to a single exchange protocol. An ML application could produce an index in the eTMF-EMS format. Any software vendor could translate this standardized index into a form that its application programming interface (API) accepts for importing documents to its proprietary database. [3]

Case Studies

Vendors offering a range of clinical and technical services are taking advantage of ML technology.

CRO Realizes Greater TMF Quality

IQVIA is leveraging OCR and ML to process documents from the trials its CRO conducts. The company plans to integrate the technology with its commercial *IQVIA eTMF* and *RIM Smart* products.

The tool identifies document types automatically, extracts relevant metadata (e.g., the language in which the document is written), supports QC, and indexes artifacts based on TMF Reference Model identifiers. Its accuracy improves through cumulative training on large repositories of trial documents.

IQVIA estimates that automated document classification and metadata extraction reduces labor by up to 50%. The company also estimates indirect savings due to improvements in quality and standardization across the TMF, resulting in less rework and fewer corrections. [4]

Accelerating TMF Service Delivery

Phlexglobal offers a range of regulatory and clinical services, including TMF management. The TMFs they manage typically hold about 700 document types. Maintaining data quality efficiently becomes challenging with manual processing at that volume.

The company employed its ML tool, *PhlexDistiller*, to prepare uploaded trial documents for creation in its proprietary eTMF,

PhlexTMF. They trained PhlexDistiller to detect common types of regulatory documents and the metadata attributes they contain. The tool presented its output to the user with suggested document indexing that the user could accept or modify. It took one-fifth of the time to prepare a document for indexing in PhlexTMF with PhlexDistiller than to index and load the same document manually.

Phlexglobal has also used PhlexDistiller to extract data from documents in support of data management software (DMS) migration and in preparing to comply with new regulatory data standards.

Return on Investment

R&D productivity can be measured as a ratio of the value of an approved therapy to the inputs invested in developing it. The timely filing of quality data is a primary input to clinical and regulatory operations. R&D organizations may ask, "Are productivity gains from implementing ML for the eTMF worth investing in now, or would it be better to wait for further improvements?"

IQVIA and Phlexglobal are employing ML tools to work with their respective eTMF and RIM products. It is likely that other competitors will take a similar approach to differentiate their product suites. Vendors could extend classification, data extraction, and indexing to applications such as clinical trial management systems (CTMSs). Costs for ML functionality may be nominal if vendors bundle it with existing software offerings.

Creating indexes in the standard eTMF-EMS format could provide additional benefits across multiple systems. A customer of an eTMF vendor offering an ML product compliant with eTMF-EMS could import documents into the eTMF and other systems regardless of vendor, provided they are also compliant with the standard. Significant manual effort is often necessary when documents are exchanged between systems. Distributing documents with standard-based indexes to multiple system APIs would remove the need for file exchanges to transfer them into the eTMF from other systems. Customers considering this use case should discuss it with their vendors. As of the date of this publication, most vendors have not yet implemented software to translate the eTMF-EMS protocol into forms their respective APIs will accept.

ML tools are already accelerating document filing and lowering manual processing costs. While future capabilities may extend these benefits, current tools can yield significant improvements in productivity. Assuming that initial costs of adopting and training ML are not excessive, the likely result is a net reduction of ongoing costs.

The overall benefits of adopting ML for the eTMF may be more difficult to quantify but are no less substantial. Improved productivity would offer opportunities to better monitor the health of trials and maintain regulatory compliance with an inspection-ready eTMF.

IMPROVING PRODUCTIVITY AND TMF INSPECTION-READINESS

Traditionally, getting documents into an eTMF has been a highly manual effort. Without the ability to automate the process of classifying and indexing documents, someone must identify metadata embedded in unstructured text and enter it manually during eTMF document creation.

Exchanging documents between systems requires someone to map the metadata terms entered by the sending party to the metadata terms the receiving party uses, which often takes many weeks to complete. ML accelerates the classification and indexing of documents. Extending its capabilities to updating the eTMF and other applications within a broad RIM framework could eliminate document file exchanges.

The case for adopting current ML technology now is strong. It may be time for more companies to rethink how they process and store regulatory documents in order to gain the most from its use.

References

[1] PhlexGlobal webinar, "Validating an AI Solution with a Focus on Clinical and Regulatory," [Online]. Available: https://www.phlexglobal.com/validating-an-ai-solution-with-a-focus-on-clinical-and-regulatory. [Accessed 12 August 2020].

[2] "Trial Master File Reference Model," [Online]. Available: https://tmfrefmodel.com/. [Accessed 18 August 2020].

[3] K. Keefer, "eTMF-EMS Business Case: Save Money And Stay Inspection-Ready With Timely TMF Content Exchange," [Online]. Available: http://keeferconsulting.com/Resources.htm. [Accessed 18 August 2020].

[4] IQVIA, "Artificial Intelligence/Machine Learning in the eTMF," in *IQVIA eTMF User Group Meeting*, Philadelphia, 2019.

[5] PhlexGlobal webinar, "How Automation is Transforming Regulatory Compliance in Life Sciences," [Online]. Available: https://www.phlexglobal.com/how-automation-is-transforming-regulatory-compliance-in-life-sciences. [Accessed 25 March 2020].

Keefer Consulting, Inc.

The goal of this whitepaper is to help clinical and regulatory operations leaders understand how machine learning could reduce the financial costs and lost opportunities associated with inefficient processes for updating the eTMF.

Keefer Consulting Inc. http://keeferconsulting.com/ helps biopharmaceutical companies improve R&D productivity and compliance through effective management of clinical and regulatory data. Ken Keefer has consulted and led teams for pharmaceutical companies and other industries for nearly 40 years. He served as project manager in the development of the eTMF Exchange Mechanism Standard, Version 1.0 for the TMF Reference Model.

To assess how your company can reduce TMF update backlogs and the risk of negative inspection findings, contact Ken Keefer at 215-462-1601 or kkeefer@keeferconsulting.com, or schedule a call at https://calendly.com/kenkeefer/15min.

The views and opinions expressed in this whitepaper are those of Keefer Consulting Inc. and should not be attributed to TMF Reference Model, IQVIA, or Phlexglobal.