PhUSE US Connect 2019

Paper ML12

Utilizing Artificial Intelligence for Efficient CRF Design

Komal Govil, Janssen R&D, Raritan, NJ, USA
Sai Vundela, Janssen R&D, Raritan, NJ, USA

ABSTRACT

Case report form (CRF) design is a time-consuming process. Many country's health authorities (HA) and institutional review boards (IRB) require draft CRFs for protocol approval long before data management is even on board. Moreover, currently there are heavy delays with creating and making updates based on current trends to standard CRFs. Likewise, creating initial CRF drafts, even with proper standards, is a very time intensive and resource consuming process. Implementing a recommender system that reduces the time delay and works more efficiently is critical to meeting the demands of future clinical trials. This paper will describe the concepts and pilot a machine learning algorithm in which information extracted from the protocol will create a preliminary list of standard global CRFs to be used for initial eCRF design.

INTRODUCTION

Efficient utilization of resources is important in the highly regulated pharmaceutical industry. CRF design is a time-consuming process, due to the technical, clinical and logistical considerations that must be made. Many country's health authorities (HA) and institutional review boards (IRB) require draft CRFs for protocol approval long before data management is even on board. Local approval of the protocol requires prompt submission of the application. Providing a stable draft CRFs to the team within time and efficient usage of the human resources is always a challenge. Automating the initial part of the CRF creation process will optimize resource utilization and allow for shorter CRF development timelines.

WHAT IS A CASE REPORT FORM?

The CRF contains data obtained during the patient's participation in the clinical trial. In clinical trials, the case report form is a way to collect relevant information required by the protocol, from each patient. The sponsor of the clinical trial aims to design the CRF to obtain only accurate and relevant information in support of the trial objectives and endpoints. A part of data management's role is to manage the logistics of CRF design and database design for data to be collected electronically at investigator sites through an electronic data capture (EDC) tool.

CURRENT CRF DESIGN PROCESS

Pre-configured standard libraries, for building electronic CRFs within EDC tools, are used to streamline data collection across trials. Standard libraries contain CRFs that can be global or disease and therapeutic area specific. The libraries usually come with preconfigured fields, simple edit checks, complex edit checks and derivations that have been validated through numerous trials. Larger sponsors tend to have large standard libraries that must be adhered to. The data management team reviews the protocol and selects the appropriate CRF's by looking into all the available libraries to see which one satisfies the requirement of the protocol. Despite the creation of these robust standard libraries, the process of selecting appropriate CRFs continues to be time-intensive, since every standard CRF must be opened separately, reviewed and then selected based on the protocol specifications.

CURRENT CHALLENGES WITH CRF DESIGN

As protocols become more complex, CRF design is also becoming more complicated. There are heavy delays with creating and making changes to standards based on the current trends. The availability of an extensive standard library has not provided a solution. More is not necessarily better. Due to the large volume of standard CRFs to be maintained, standards are slow to adapt. Therefore, the CRF designer is usually working with outdated standards and repeating the mistakes that have been resolved on more recent trials. For example, if there is a design flaw in a CRF that was resolved in a recent study, but has not been incorporated in organization standards, it will continue to propagate and appear in other studies. In a similar instance, the FDA recently mandated the removal of the date of birth field from the CRF. The draft CRF submitted to the local health authorities, should reflect the latest standard, however, it might take large organization time to update their standard forms and it might delay the overall

submission process. Remember, some IRBs and HAs meet infrequently, and this can have a huge impact on the overall timelines.

Moreover, selecting the applicable standard CRF is a time consuming for CRF designers who must be able to distinguish the subtle differences between similar standard CRFs. For instance, there are many study drug administration forms available in the library. In the current process, the designer finds the object identifier (OID) of all exposure forms. Each pdf file is opened individually, and manually reviewed. The designer then selects the most suitable form(s) for the protocol. Because of the minor differences in the forms, there is a greater chance of selecting an incorrect form.

WHAT IS ARTIFICIAL INTELLIGENCE

Artificial Intelligence (AI) is around us everywhere in computers today. Siri can recognize our speech commands, Facebook is able to recognize and tag our friends in photos. Google translate can help us read signs in other languages. These are all examples of how man-made, artificial computers can perform tasks that are typically done by humans (intelligence). AI is not automation, in which explicit rules are given to the system to carry out. Automation works well for repetitive tasks, while artificial intelligence works to generalize tasks so that a machine can learn them. Machine learning is an application of AI, in which conclusions are learned from data by computer algorithms. It is magic. The pharmaceutical industry is actively exploring the use of artificial intelligence for drug design, predicting treatment results, protocol development, drug dosage and many other areas.

RECOMMENDER SYSTEMS

A recommender system is an information filtering algorithm that can be used to recommend products based on unique individual preferences. The two most common types of such systems are collaborative filtering and content-based filtering. A collaborative filter considers the behavior of other similar users and recommends products based on their behavior. For example, social media, such as Facebook and LinkedIn, suggest friends that your friends have friended. Content based systems recommend items based on latent commonalities between items. For example, if you like a Nikki Minaj song on Pandora, it will play other songs by Nikki Minaj on that station.

Netflix uses a hybrid system combining the two approaches to create a more precise recommender system. For CRF design such a hybrid system would also be useful. The collaborative portion of the system would help CRF design adapt quickly to current trends and consider similar studies. The content-based portion of the algorithm would help identify similarities between the forms and the protocols based on features such as therapeutic areas, compound specifications and the trial phases.

OBJECTIVE

Implementing a system that reduces time and works more efficiently is a necessity to meet the demands of future clinical trials. The objective of this paper is to assess the proposal of developing a machine learning algorithm that can take a protocol and extract the necessary information to create a preliminary list of standard CRFs to be used as an initial CRF draft. If we can effectively utilize the technology and adopt that to our needs many of the existing design challenges can have a solution.

DISCLAIMER

The scope of this paper is to present the opinions and suggestions of the authors. The interpretations of standards and procedures contained in this paper are those of the authors and are not necessarily correct. Any views and recommendations stated within this paper are those of the authors, and they do not represent the positions of their employer.

THE APPROACH

USER AND ITEM PROFILES: A HYBRID APPROACH

A user profile is the set of features that can be used to describe the user. Similarly, the item profile is the set of features that can be used to describe the item. In this case, each protocol is a user and each CRF is an item. The end goal of the program is to provide us with the set of standard form OIDs that are most applicable to the protocol, as well as a pdf combining all unique forms selected. As mentioned above, we will be following a hybrid approach. The content-based recommendation system will compare the user profile to the item profile to find the best recommendations to include. The collaborative portion of the recommender system will compare the user profile to the similar recent protocols. The hybrid system will help resolve the cold start problem many recommender systems have. The cold start problem is when a new user is introduced but has trouble being matched since there is no information on that user.

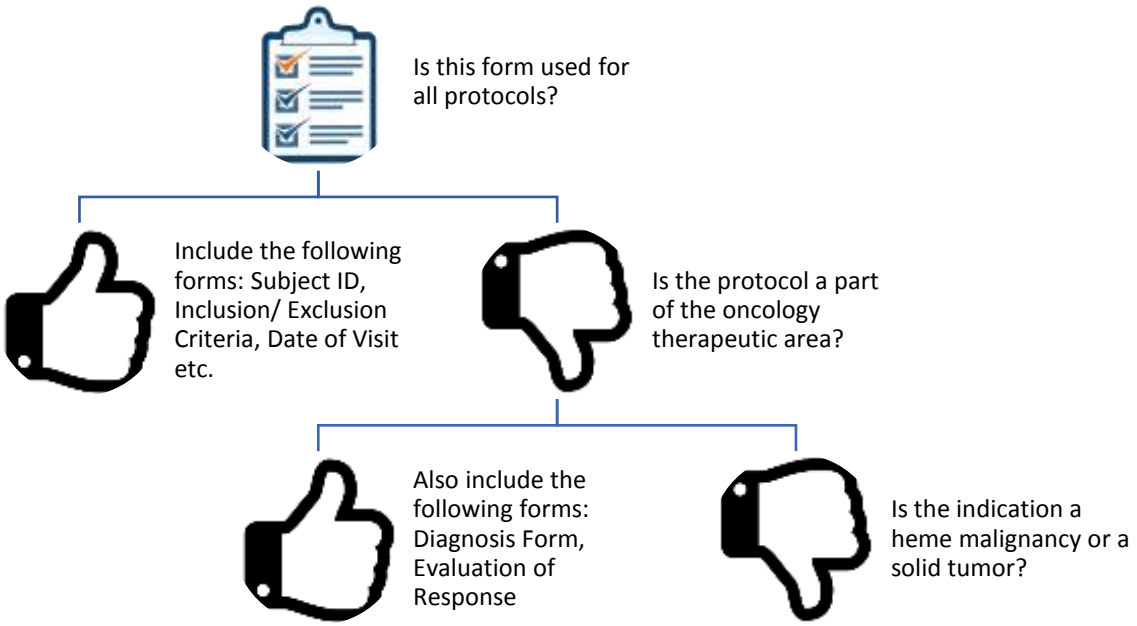
TRAINING AND TESTING THE MODEL

A proof of concept will be performed to test the theory. We have selected six oncology protocols that followed a full governance model. This means that all CRFs used in these protocols were from standards and no new CRFs were created specifically for the protocol. We will randomly take four protocols to be our train dataset and train the machine learning algorithm to create a model based on these protocols. The remaining two protocols will be used to test our algorithm.

The test dataset is to confirm the accuracy of the machine learning model and prevent overfitting to the train protocols. A program could output the perfect CRF for the protocols used to train it. If it cannot be generalized to other new protocols, then that program would be useless despite the perfect match. Similarly, in this case a false negative is much more dangerous than a false positive. In the case of a false positive the CRF reviewers would simply remove the extraneous CRF, however in the case of a false negative, they may miss required fields and impact the study.

DECISION TREES

For this classification problem we will be using a machine learning decision tree classifier. Many of us have played the game 20 questions on a long car ride. In this game we can ask up to 20 'yes' or 'no' questions to figure out what the other player is thinking. A guesser can use up their 20 questions asking if the thinker is thinking of very specific items, like a watch, or a computer. A better strategy would be to try and narrow down the category of the object. The program will follow a similar approach. The program can order and find the minimum number of questions that are required to find a solution, without explicit rules.



FEATURE SELECTION

While designing the program, a deep expertise in clinical data management is required to select the most appropriate features for this task. Determining which features provide the most accurate representation of each item and user will be the most challenging part of this endeavor. The end user will benefit greatly, as the algorithm will find the crucial features that can represent items and users. As a part of feature engineering, new features can also be created from existing features as needed.

NATURAL LANGUAGE PROCESSING

While a computer can parse programming languages, natural languages provide unique challenges. For example, the words 'to eat', 'eating', 'eats', 'ate' have a similar semantic meaning. While humans can easily pick this up, a computer will struggle. One way to deal with this problem is lemmatization, which basically reduces a word down to its essence by removing prefixes and suffixes. Similarly, a human can tell that 'cancer', 'lesions', 'tumors', 'leukemia' and 'lymphoma' will indicate a therapeutic area of oncology. Tokenization, the combination of words into one concept,

can help us resolve this issue. Another problem is the use of stop words, such as 'a', 'an', 'the' and, 'of' in the English language. There are many programs that can help remove these stop words. Large text is difficult to search and to store, therefore shortening the text to its essential concepts is necessary. Term frequency-inverse document frequency (TF.IDF) calculates the weight of how often a word is used in a corpus. The most common words can then be used to describe the text. Many methods can be used to characterize the essence of the text and give a set of features to describe the protocol. The challenge is determining the best set of features that can give us the most accurate algorithm.

DIMENSIONALITY REDUCTION

Theoretically, we could store each word of the protocol as a feature of the protocol. But this would result in garbage in and garbage out. Not our goal. Similarly, too many features will yield large unwieldy datasets that are simply unnecessary to extract the needed information. Dimensionality reduction techniques, such as principal component analysis (PCA), can be used to help determine which features aid the algorithm and which features are unnecessary. PCA can be used to analyze which features of a profile provide the most useful information to the algorithm. Features that do not provide a lot of information necessary to make predictions can then be eliminated. PCA is a method of summarizing data and removing redundant features while maintaining the variations in the original dataset. Many such methods are available in python libraries.

PYTHON LIBRARIES

Python is a versatile and simple computer language with many machine learning extensions. The following open source libraries can be imported to begin building an algorithm. The python data analysis library (pandas) provides high-level data structures that are intuitive and easy to use. It also has inbuilt methods for grouping, combining and filtering data. Numpy is a dependency of pandas and is a library used for scientific computing. Scikit learn is a machine learning library that contains algorithms for tasks such as clustering, regression and classification. Natural language toolkit (NLTK) provides text classification libraries for processing natural language.

```
import pandas as pd
import numpy as np
import sklearn
import nltk
```

COST BENEFIT ANALYSIS

Manual intervention is still required for quality control. After receiving an initial draft from standard CRF forms, the CRF designer will still need to review all forms to ensure protocol adherence. They will still also need to design new protocol specific forms which may or may not get added to standards in the future. The higher the number of new forms, the more time consuming the design process will still be.

Creating such an algorithm will require an initial investment of time and resources. The benefits will only be felt in the long run. There may be an initial resistance to another new platform that everyone will need to learn.

Still, the end benefits will outweigh the initial discomfort. The new process will be cost efficient and will significantly shorten CRF design time. It will implement a system that can adapt to change by learning from other trials and help maintain consistency between similar trials. It will also eliminate many unproductive tasks such as combining pdfs from the CRF designers daily work. That time and energy can then be focused towards better design for new CRFs.

CONCLUSION

Focusing on a limited number of protocols and on the global library, will aid implementation in the beginning. Of course, for machine learning algorithms more data is better, and we can expand the program to take in more protocols. If the proof of concept is successful, eventually it can easily be applied to therapeutic area and compound specific standard eCRFs. CDASH standards can be included as well as company specific standards. It may even be implemented on a field level rather than a form level. Furthermore, it can help streamline the process of creating CRF standards. Moreover, other such processes may benefit from such powerful technology.

REFERENCES

Bobadilla, J., Ortega, F., Hernando, A., Gutiérrez, A. (2013). Recommender Systems Survey. *Knowledge-Based Systems, 46,* 109-132.

[Google Developers]. (2016, March 30). Hello World – Machine Learning Recipes # 1 [Video File]. Retrieved from https://www.youtube.com/watch?v=cKxRvEZd3Mw&list=PLOU2XLYxmsIIuiBfYad6rFYQU_jL2ryal

Iriondo, R. (2018, October 15). Differences Between AI and Machine Learning and Why it Matters. Retrieved January 11, 2019, from https://medium.com/datadriveninvestor/differences-between-ai-and-machine-learning-and-why-it-matters-1255b182fc6

Mwiti, D. (2018, October 03). How to build a Simple Recommender System in Python – Towards Data Science. Retrieved January 11, 2019, from https://towardsdatascience.com/how-to-build-a-simple-recommender-system-in-python-375093c3fb7d

Park, D.H., Kim, H.K., Choi, I.Y., & Kim, J.K. (2012). A literature review and classification of recommender systems research. *Expert Syst. Appl., 39,* 10059-10072.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Author Names: Komal Govil and Sai Vundela
Company: Janssen Research and Development
Address: 920 US 202 South
City / Postcode: Raritan, NJ 08869
Work Phone: 908 218 7622, 908 927 4940
Email: kgovil@its.jnj.com; svundela@its.jnj.com