A B S T R A C T

The success of clinical trial operations depends heavily on enroling patients on time. Industry wide, most trials experience delays and fail to enrol as initially planned. In the first part of this thesis, we have developed a data science model to predict the enrolment speed of a clinical trial. With this, enrolment simulations and predictions can be run at design stage. This model is based on more than 3500 historical trials, more than 100 relevant features and three outcome variables, i. e., (1) the appropriate duration (in months) to recruit 50 patients, (2) the enrolment duration (in months) of patient recruitment given a target size and (3) the number of patients per site per month enroled. We implemented and evaluated four predictive models, i. e., the LASSO Linear Regression model, Random Forest Regression model, XGBoost Regression model and AutoML Regression model, using two performance metrics, i. e., the MSE and adjusted R-squared. Based on the simplicity, explicability and prediction outcomes, the XGBoost Regression model performs the best with adjusted R-squared values of 0.76 for outcome variable (1) the appropriate duration (in months) to recruit 50 patients; 0.43 for outcome variable (2) the enrolment duration (in months) of patient recruitment given a target size; and 0.55 for outcome variable (3) the number of patients per site per month enroled. Regarding the second part of this thesis, we found that the Bayes method is a very appropriate and promising mathematical approach for reprojecting timelines with interval estimates to predict as accurately as possible the Last-Patient-In (LPI) time in the course of patient recruitment for a clinical trial, given a target number of patients. This takes into account the uncertainties inherent in the estimated parameters from the statistics and in the random process of the time data in predicting accruals. In addition, this method uses prior information based on quantifiable information, which make this prediction to be fairly objective and consensus-based.

In overall, the work in this thesis can allow for more data-driven solutions to strengthen both enrolment speed and LPI predictions for clinical trials.

Z U S A M M E N FA S S U N G

Der Erfolg klinischer Studien hängt in hohem Maße davon ab, dass die Patienten rechtzeitig eingeschrieben werden. In der gesamten Branche kommt es bei den meisten Studien zu Verzögerungen, und die Patienten werden nicht wie ursprünglich geplant aufgenommen. Im ersten Teil dieser Arbeit haben wir ein datenwissenschaftliches Modell zur Vorhersage der Einschreibegeschwindigkeit einer klinischen Studie entwickelt. Mit diesem Modell können bereits in der Planungsphase Simulationen und Vorhersagen zur Rekrutierung durchgeführt werden. Dieses Modell basiert auf mehr als 3500 historischen Studien, mehr als 100 relevanten Merkmalen und drei Ergebnisvariablen, diese sind, (1) die angemessene Dauer (in Monaten) für die Rekrutierung von 50 Patienten, (2) die Rekrutierungsdauer (in Monaten) der Patientenrekrutierung bei einer gegebenen Zielgröße und (3) die Anzahl der Patienten pro Standort und Monat. Wir implementierten und bewerteten vier prädiktive Modelle, diese sind, das lineare Regressionsmodell LASSO, das Regressionsmodell Random Forest, das Regressionsmodell XGBoost und das Regressionsmodell AutoML, unter Verwendung von zwei Leistungsmetriken, diese sind, MSE und adjustiertes R-Quadrat. Auf der Grundlage der Einfachheit, der Erklärbarkeit und der Vorhersageergebnisse schneidet das XGBoost-Regressionsmodell mit einem adjustierten R-Quadratwert von 0,76 für die Ergebnisvariable (1) die angemessene Dauer (in Monaten) der Rekrutierung von 50 Patienten; 0,43 für die Ergebnisvariable (2) die Dauer (in Monaten) der Patientenrekrutierung bei einer gegebenen Zielgröße; und 0,55 für die Ergebnisvariable (3) die Anzahl der Patienten pro Standort und Monat am besten ab. In Bezug auf den zweiten Teil dieser Arbeit haben wir festgestellt, dass die baysianische Methode ein sehr geeigneter und vielversprechender mathematischer Ansatz für die Reprojektion von Zeitreihen mit Intervallschätzungen ist, um die Last-Patient-In (LPI)-Zeit im Verlauf der Patientenrekrutierung für eine klinische Studie bei einer gegebenen Zielanzahl von Patienten so genau wie möglich vorherzusagen. Dabei werden die Unsicherheiten berücksichtigt, die in den geschätzten Parametern aus der Statistik und im Zufallsprozess der Zeitdaten bei der Vorhersage der Rekrutierung liegen. Darüber hinaus verwendet diese Methode Prior-Informationen, die auf quantifizierbaren Informationen beruhen, wodurch diese Vorhersage ziemlich objektiv und konsensbasiert wird.

Insgesamt kann die Arbeit in dieser Thesis mehr datengesteuerte Lösungen ermöglichen, um sowohl die Einschreibegeschwindigkeit als auch die LPI-Vorhersagen für klinische Studien zu verbessern.