Question 3

The rise of machine learning (ML) models has improved efficiency and productivity in advancing research and innovation. However, a major concern in the development and practice of ML models is the presence of data doppelgangers, within training and validation datasets that are highly similar to each other due to chance or other factors (Wang et al., 2022). This phenomenon could yield unreliable validation results and hinder the accuracy of ML models. This report aims to examine whether data doppelganger effects are unique to biomedical data, propose ways to avoid or check for such effects in the development of ML models in medicine and pharmacology.

In biomedical data, doppelganger effects arise when the training and validation datasets are highly similar in terms of patient demographics, disease prevalence, etc. For example, a study on the prediction of myocardial infarction risk using electronic health records found that the use of highly similar training and validation datasets led to overfitting and inflated performance metrics (Xiong et al., 2022). Similarly, a study on hospital mortality using clinical data also found that the use of data doppelgangers led to unreliable model performance (Seki et al., 2021).

Data doppelganger effects are not unique to biomedical data, other data types like that of imaging data also face challenges when the training and validation datasets are highly similar in terms of image acquisition parameters, such as scanner type, resolution, and field of view (Varoquaux & Cheplygina, 2022). Classification of tumorous masses found within the brain using magnetic resonance imaging (MRI) found that ML models overfit the parameters resulting in poor generalisation of validation results (Powell et al., 2016).

To avoid doppelganger effects in biomedical data, it is crucial to systematically evaluate the quality and diversity of training and validation datasets. This can be achieved by using independent datasets derived from different sources or time periods, stratifying the datasets based on relevant variables, and using cross-validation techniques to assess model performance (Wang et al., 2022). Other studies suggest the use of lock boxes or blind analyses after the model's hyperparameters have been determined (Powell et al., 2016).

Overall, data doppelgangers are difficult to detect in large data sets. Doppelganger effects deter the development and use of effective and accurate machine learning models. However, by carefully selecting and pre-processing the data, using appropriate validation techniques, mitigation is possible in order to develop models that are robust, reliable, and effective.

References

Powell, M. et al. (2016) "I tried a bunch of things: The dangers of unexpected overfitting in classification." Available at: https://doi.org/10.1101/078816.

Seki, T., Kawazoe, Y. and Ohe, K. (2021) "Machine learning-based prediction of in-hospital mortality using admission laboratory data: A retrospective, single-site study using Electronic Health Record Data," PLOS ONE, 16(2). Available at: https://doi.org/10.1371/journal.pone.0246640.

Varoquaux, G. and Cheplygina, V. (2022) "Machine Learning for Medical Imaging: Methodological Failures and recommendations for the future," npj Digital Medicine, 5(1). Available at: https://doi.org/10.1038/s41746-022-00592-y.

Wang, L.R., Wong, L. and Goh, W.W. (2022) “How doppelgänger effects in biomedical data confound machine learning,” Drug Discovery Today, 27(3), pp. 678–685. Available at: https://doi.org/10.1016/j.drudis.2021.10.017.

Xiong, P., Lee, S.M.-Y. and Chan, G. (2022) “Deep learning for detecting and locating myocardial infarction by electrocardiogram: A literature review,” Frontiers in Cardiovascular Medicine, 9. Available at: https://doi.org/10.3389/fcvm.2022.860032.