Integrating clinical data from cross-sectional and longitudinal studies

Yuanxi Li & Allan Tucker
Brunel University, London. United Kingdom
{yuanxi.li, allan.tucker}@brunel.ac.uk

***Abstract*— Clinical trials are typically conducted over a population in order to illuminate certain characteristics of a health issue or disease process. These cross-sectional studies provide a snapshot of these disease processes over a large population but do not allow us to model the temporal nature of disease. Longitudinal studies on the other hand, are used to explore how these processes develop over time but can be expensive and time-consuming, and only cover a relatively small window within the disease process. This paper explores a technique for integrating cross-sectional and longitudinal studies to build models of disease progression.**

Keywords- cross-section, data integration, disease progression

I. INTRODUCTION

Degenerative diseases such as cancer, Parkinson's disease, and glaucoma are characterised by a continuing deterioration to organs or tissues over time. Longitudinal studies [1] measure clinical variables from a number of people over time generating Multivariate Time-Series (MTS) data. The advantage of longitudinal data is that temporal details of the disease progression can be determined. However, data is often limited in terms of the cohort size, due to the expensive nature of the studies. Cross-sectional studies record attributes (such as clinical test results) across a sample of the population, thus providing a snapshot of a particular process [2]. An advantage of cross sectional studies is that they capture the diversity of a sample of the population and therefore the degree of variation in the symptoms. They are also relatively cheap compared to longitudinal studies that involve extensive follow up. The main disadvantage of such studies is that the progression of disease is inherently temporal in nature and the time dimension is not captured. Previously, we developed a resampling approach known as the temporal bootstrap [3] that builds multiple trajectories through cross sectional data to approximate genuine longitudinal data. These pseudo time-series can be used to build approximate temporal models for prediction and for identifying stages in disease progression [4]. However, the use of cross-sectional data to build these models will always be limited by the fact that no genuine timestamps have been used to infer the models and so only an ordering is captured.

II. METHODS

Here, we investigate the effect of incorporating longitudinal data into pseudo temporal models in order to calibrate them. We explore how to best balance cross-sectional data and longitudinal data in order to minimise the expensive process of longitudinal data collection. Pseudo Time-Series (PTS) can be used to build temporal models such as Hidden Markov Models for forecasting [5]. The Temporal BootStrap (TBS) builds PTS by resampling data from a cross-sectional study and repeatedly building trajectories through the samples. Each trajectory begins at a randomly selected datum from a healthy individual and ends at a random datum classified as diseased. The trajectory is determined by the Floyd-Warshall algorithm [6], a well-established algorithm for finding the shortest path in weighted graphs. A full description of the algorithm to generate PTS appears in [3] and example PTS generated from simulated cross-sectional data are shown in Figure 2.

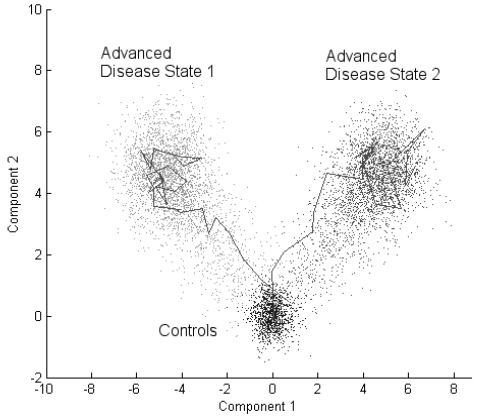

Figure 1. Example PTS generated from TBS on Simulated Data

We explore whether adding a small number of longitudinal data samples to models learnt from cross-sectional data (via the PTS approach) improves them. Real data from 91 Visual field tests where patients who are at high-risk of developing glaucoma undertake a psychophysical test to identify damage to sectors of their vision. No gold standard model exists but a comparison can be made to models learnt on the time-series and on sampled cross-sections of the time-series: We sample one VF test from each of the 91 patient time-series (91 MTS VF DATA in Figure 4) to generate a cross-sectional sample and generate PTS data for learning models from (PTS). We use AutoRegressive HMMs (ARHMMs) to model the data as we found it captures the smooth progression of disease. We compare this model as well as ones learnt from a combination of PTS and 10/20 real time-series (Random 10/20 MTS) to see how quickly we can learn models that are close to the original. This is achieved by comparing the Kulbaeck Leibler (KL) distances [6] between these calibrated models and the mean KL distance between 200 different ARHMMs learnt from the same original time-series (MEAN VARIANCE). In other words, if we can learn models from the sampled CS data that have similar KL distances to the general variation in learning a model from the full time-series, then we assume that the models are as

close to one learnt from a full time-series. The experiments are repeated 100 times to derive confidence intervals and Wilcoxon Rank statistics on the distances.

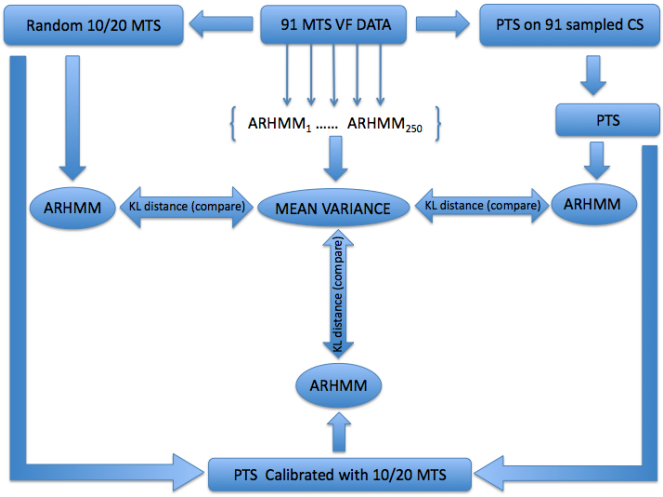

Figure 2. Figure of VF Data Experimental Framework

III. RESULTS

Figure 3 shows the KL distributions generated from the experiments. Notice firstly that the KL distance between models that have been learnt on the full 91 time-series are in the region of 80-90 with a small confidence interval denoting relatively small variance from one model learning to the next. The models that are learnt from the sampled cross-section using the PTS approach are impressively close to the time-series models but distinctly higher in KL distance (likely to be because real temporal information is lacking).

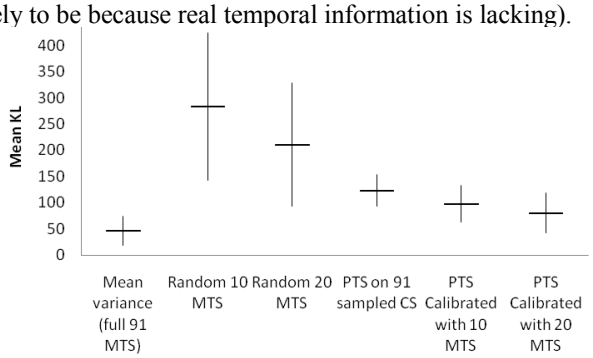

Figure 3. KL results for VF data with confidence intervals

When 10 and 20 real time-series are used to calibrate the model, however, we see further improvement in the KL distance resulting in models that are demonstrably closer to the models learnt from all 91 time-series. Finally, models that are learnt from using the relatively small number of calibrating time-series only are clearly worse with much higher distance and large confidence intervals. Looking at Wilcoxon rank tests for significance in Table I, nearly all models are indeed significantly worse than the variation between models learnt on the full longitudinal dataset (significant differences are marked with asterisks) except for the PTS model calibrated with 20 real time-series. This implies we can learn models that are as good as the natural variation between model building on full longitudinal data by building PTS and calibrating with only 20 real longitudinal samples. We also see that many of the inferior models are similar in terms of their distances except for the very worst models (learnt from only 10 time-series). These are different (worse) from the calibrated PTS models.

IV. CONCLUSIONS

In this paper we have explored to what degree pseudo time-series, learnt from building trajectories through a cross-sectional study, can be "calibrated" by relatively small numbers of real longitudinal study data to gain the advantage of both types of study. Future work will involve exploring Bayesian approaches to integrate longitudinal and cross-sectional data where cross sectional data is used to build a prior model which can be updated with longitudinal data.

REFERENCES

[1] Albert, PS. Longitudinal data analysis (repeated measures) in clinical trials. Statistics in medicine, 18(13):1707–1732, 1999.

[2] Mann, CJ. Observational research methods. research design ii: cohort, cross sectional, and case-control studies. Emergency Medicine Journal, 20(1):54–60, 2003.

[3] Tucker, A. Garway-Heath, D. The pseudo temporal bootstrap for predicting glaucoma from cross-sectional visual field data, IEEE Trans IT Biomed, 14 (1) (2010), pp. 79–85

[4] Li, Y., Swift, S. and Tucker, A., Modelling and analysing the dynamics of disease progression from cross-sectional studies, Journal of Biomedical Informatics 46 (2) : 266- 274, 2013

[5] Rabiner, LR. A tutorial on hidden Markov models and selected applications in speech recognition. Proc. of the IEEE, 257-286, 1989.

[6] Floyd. RW. Algorithm 97: shortest path. Communications of the ACM, 5(6):345, 1962.

[7] Kasza, J. and Solomon. PJ. A comparison of score-based methods for estimating Bayesian networks using the Kullback Leibler divergence. arXiv:1009.1463v2 [stat.ME]. In press for Communications in Statistics: Theory and Methods, 2013.

TABLE I. WILCOXON RANK SIGNIFICANCE (SIGNIFICANT P VALUES ARE MARKED WITH AN ASTERISK P<0.01)

	Mean variance	Rand 10	Rand 20	PTS	PTS Cal(10)	PTS Cal(20)
Mean variance (full 91 MTS)	-	**0.000***	**0.001***	**0.001***	**0.005***	0.011
Random 10 MTS	-	-	0.975	0.023	**0.002***	**0.001***
Random 20 MTS	-	-	-	0.042	0.014	0.010
PTS on 91 sampled CS	-	-	-	-	0.452	0.327
PTS Calibrated with 10 MTS	-	-	-	-	-	0.773
PTS Calibrated with 20 MTS	-	-	-	-	-	-