Informatics for Health: Connected Citizen-Led Wellness and Population Health
R. Randell et al. (Eds.)

doi:10.3233/978-1-61499-753-5-261

Informative Observation in Health Data: Association of Past Level and Trend with Time to Next Measurement

Matthew SPERRIN [a,1] , Emily PETHERICK [b] and Ellena BADRICK [a]
[a] *Farr Institute, Faculty of Biology, Medicine and Health, University of Manchester, Manchester Academic Health Science Centre*
[b] *School of Sport, Exercise and Health Sciences, Loughborough University*

Abstract. In routine health data, risk factors and biomarkers are typically measured irregularly in time, with the frequency of their measurement depending on a range of factors – for example, sicker patients are measured more often. This is termed *informative observation*. Failure to account for this in subsequent modelling can lead to bias. Here, we illustrate this issue using body mass index measurements taken on patients with type 2 diabetes in Salford, UK. We modelled the observation process (time to next measurement) as a recurrent event Cox model, and studied whether previous measurements in BMI, and trends in the BMI, were associated with changes in the frequency of measurement. Interestingly, we found that increasing BMI led to a lower propensity for future measurements. More broadly, this illustrates the need and opportunity to develop and apply models that account for, and exploit, informative observation.

Keywords. Informative observation, Longitudinal modelling, Observation processes.

1. Introduction

When conducting longitudinal statistical analysis with routinely collected health data, it is often assumed that the process that governs whether and when data are observed – the observation process – is *ignorable*. This statistically defined term means that we do not need to concern ourselves directly with the observation process, and it is not necessary to model the process explicitly. In real terms this translates to assuming that measurements of a risk factor or biomarker are regularly spaced (e.g. measured once per year), or that they are irregularly spaced but the spacing is not informative (conditional on measured covariates). There is an approximate correspondence with the related concepts in missing data of missing completely at random and missing at random.

Intuitively, however, observations are made according to an underlying process driven by the patient, the clinician, and the environment. Therefore, the timing of observations may be informative, over and above the actual values observed (again, this corresponds to missingness not at random). For example, a patient concerned about

[1] Corresponding author, Vaughan House, University of Manchester, M13 9GB;
E-mail: matthew.sperrin@manchester.ac.uk.

their health may engage with the health service more and hence have smaller gaps between measurements. A clinician concerned about the health of a patient may request to see them again sooner. The propensity for a subsequent observation of a risk factor may also depend on previous observed values, and trends in previous values of the risk factor. For example, if a biomarker is rapidly rising, the clinician may wish to measure it again within a short time period. This is termed *outcome-dependent follow-up*. Some methods to handle data subject to a non-ignorable observation process are available in the statistical literature [1], [2]; these are based on assuming a joint model for both the observation process and the outcome process. However, these have seen limited application to routinely collected health data. Moreover, existing approaches typically view the observation process as a 'nuisance' and not to be of scientific interest [3]; we hypothesise that the observation process can be exploited to gain additional information for inference.

In this paper we explored the properties of the observation process in the example of body mass index (BMI) measures for patients with type 2 diabetes (T2D), taken in primary care in Salford, UK. We hypothesised that BMI measurements would depend not only on patient demographics but also on previous measurements, and the current trend, of BMI – i.e. outcome-dependent follow-up.

2. Methods

We used anonymized primary care data from the Salford Integrated Record. Salford, UK, is a relatively deprived city in Greater Manchester, UK, with a population of approximately 300,000, served by a single hospital and 53 GP practices. Our study period was 1 April 2004 to 31 December 2012. The start date was chosen to align with the quality and outcomes framework (QOF), which is a scheme, started in 2004, in which GPs are incentivized to meet a range of indicators that promote patient care; one of these indicators is that T2D patients have a BMI measurement within the financial year. Individuals were considered 'at risk' for a BMI measurement during this period provided that they had received a T2D diagnosis and were alive. BMI measurements outside of this time range were excluded; however, they were used where appropriate as 'previous BMI' readings. An individual may have no BMI readings recorded at all, but still be included in the analysis (since they are still 'at risk' of a BMI measurement). If an explicit T2D diagnosis date was not available, the date of first prescription of anti-diabetic medication was used as a proxy for this. If neither of these were available the patient was removed from the analysis. Patients were also removed if no date of birth was available. Finally, patients who were younger than 35 or older than 85 at diagnosis date were also removed.

We built a statistical model that focused on the observation process (times at which BMI is observed) rather than the outcome itself (the BMI measurements). Specifically, we used a Cox proportional hazards model for recurrent events to model time to next BMI measurement. We used age as the timescale, and left truncated at the study start date or diabetes diagnosis date, whichever was later. The earliest of death and the study end date was considered a right censoring event. The model incorporated frailty terms to capture within-person correlation [4].

In our multivariable model, covariates underlying the observation process of primary interest were: the previous BMI measure; the difference between the previous measure and the one before, which represents a trend that would be observable by the

GP; and the difference between the current (yet to be taken) BMI measure and the previous one, which may reflect the patient's current perception of weight change. Time since diagnosis and calendar year were included as time-updated terms. We also included gender, and separate indicator variables for diagnosis of coronary heart disease (CHD), chronic obstructive pulmonary disease (COPD), asthma and cancer.

All analyses were carried out using R version 3.2.0 [5].

3. Results

A data exclusion flow chart, both at the patient level and individual BMI observation level, is given in Figure 1; the final dataset comprised 11,805 patients with a total of 133,425 BMI readings.

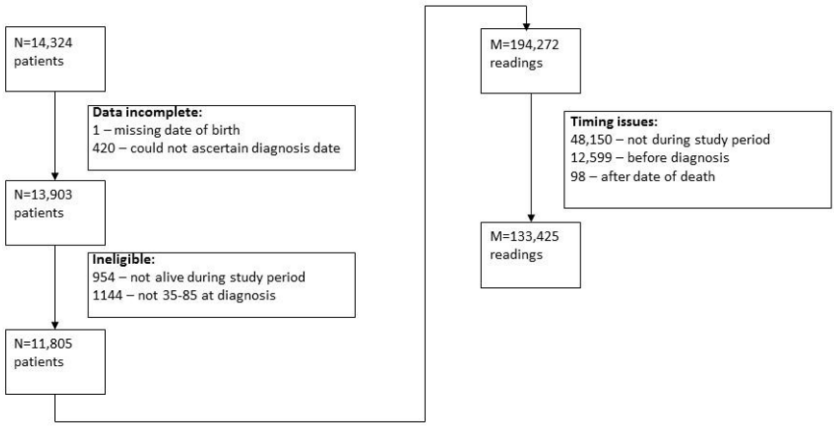

Figure 1. Data exclusion flowchart. Ineligible patients are excluded first, then ineligible readings are removed for eligible patients.

Baseline information is given in Table 1. The mean first BMI was 31.28 (in the obese category, which is as expected for T2D patients), and we observed a median of 9 BMI measurements per patient.

Table 1. Baseline information. (SD = standard deviation; IQR = interquartile range).

BMI at baseline (first reading in time period)	Mean = 31.28, SD = 6.41
Number of BMI measurements	Median = 9, Min=0, Max=112, IQR = 10
Year of birth	Median = 1944, Min = 1911, Max = 1976
Age at baseline	Mean = 62.02, SD = 11.74
Male	N = 6647 (56.31%)
CHD (ever)	N = 4183 (35.44%)
COPD (ever)	N = 1744 (14.77%)
Asthma (ever)	N = 2268 (19.21%)
Cancer (ever)	N = 1463 (12.39%)
Dead before 31/12/2012	N = 539 (4.57%)

Table 2 gives the proportion of patients for whom at least one BMI measure is made within a financial year (out of all patients who are alive and have a T2D

diagnosis for the entire financial year). This proportion increases steadily from 0.739 to 0.825 across the study period.

Table 2. Proportion of patients for whom at least one BMI measure is made within a QOF year (out of all patients who are alive and have a T2D diagnosis for the entire QOF year).

Year	Proportion with BMI reading	Number eligible
2004-5	0.739	6345
2005-6	0.752	6962
2006-7	0.780	7659
2007-8	0.806	8292
2008-9	0.815	8958
2009-10	0.826	9383
2010-11	0.834	9776
2011-12	0.825	10039

Results of the recurrent event proportional hazards model are given in Table 3, with hazard ratios given per unit BMI or per year as appropriate. We see that propensity (or hazard) to measure BMI was increased if the previous BMI reading was higher, but decreased if there was an observed or perceived upward trend in BMI. The difference between the two previous BMI readings had a larger effect on the hazard than the difference between the current and previous readings. CHD, COPD and asthma patients all had a higher propensity/hazard to be measured, while for cancer there was no significant difference in the hazard. There was no evidence of a violation of the proportional hazards assumption (P=0.957 in global test, and no individual covariates had significant relationship between Schoenfeld residuals and time).

Table 3. Hazard ratios from Cox recurrent event model.

Variable	HR (95% CI)
Previous BMI	1.010 (1.008,1.012)
Difference between previous reading and reading before	0.979 (0.976,0.983)
Difference between current and previous reading	0.985 (0.982,0.988)
Male (reference: female)	0.996 (0.973,1.021)
Calendar year	0.954 (0.951,0.957)
Time since diagnosis	1.027 (1.025,1.030)
CHD presence	1.026 (1.001,1.052)
COPD presence	1.127 (1.089,1.165)
Asthma presence	1.057 (1.024,1.090)
Cancer presence	0.983 (0.948,1.018)

4. Discussion

Contrary to our prior expectation, an increasing trend in BMI lowered the propensity for a repeat measurement of BMI. This is surprising and potentially concerning, and needs to be understood clinically. It was reassuring to find no evidence of a gender difference.

The findings show that the propensity to measure BMI depends on previous measurements and trends. This is likely to hold for other measures such as blood pressure and cholesterol. Appropriate statistical modelling techniques need to be used to account for this outcome dependent, non-ignorable structure in the observation

process, to prevent biased inference. One such approach is to build joint models for the observation process and outcome process (e.g. [6]). Mixed effect models can also be applied with limited bias for estimation of fixed effects [7]; however estimation of random effects can be badly biased [3].

Rather than viewing informative observation as a nuisance, we suggest that the presence of observations can be used for prediction. For example, we have shown that engagement with smart weighing scales (i.e. presence of weight measurements) is an independent predictor for weight loss [8].

The main strength of the study is that we use sophisticated modelling techniques to understand the observation process, which is typically ignored in the literature. A limitation is that we only considered BMI, although we expect that the findings will generalise to other clinical risk factors and biomarkers that may be measured irregularly over a patient's life course. Modelling limitations include that we have used ever/never terms for other diseases and smoking status – a time dependent approach could also have been considered. We could in theory also have incorporated other variables like blood pressure into the model; however they are themselves subject to irregular and potentially outcome-dependent follow-up. Moreover, a number of other variables were excluded that could explain changes in BMI – particularly T2D treatments such as metformin. We took a pragmatic approach to variable inclusion in this paper as we sought only to demonstrate the concept.

This study has shown in a real example that the observation process of a clinical risk factor may depend on previous measurements. It may also depend on other factors, measured or unmeasured. This is an area that brings challenge and opportunity: the challenge to produce models that are not biased by the presence of informative observation, and the opportunity to use the observation process itself in prediction.

References

[1] H. Lin, D. O. Scharfstein, and R. A. Rosenheck, "Analysis of longitudinal data with irregular, outcome-dependent follow-up," J. R. Stat. Soc. Ser. B (Statistical Methodol., vol. 66, no. 3, pp. 791–813, 2004.

[2] J. Sun, D.-H. Park, L. Sun, and X. Zhao, "Semiparametric Regression Analysis of Longitudinal Data With Informative Observation Times," J. Am. Stat. Assoc., vol. 100, no. 471, pp. 882–889, Sep. 2005.

[3] C. E. McCulloch, J. M. Neuhaus, and R. L. Olin, "Biased and unbiased estimation in longitudinal studies with informative visit processes," Biometrics, vol. 72, no. 4, pp. 1315–1324, Dec. 2016.

[4] C. McGilchrist and C. Aisbett, "Regression with frailty in survival analysis," Biometrics, 1991.

[5] R. Team, "R Development Core Team," R A Lang. Environ. Stat. Comput., 2013.

[6] J. Sun, L. Sun, and D. Liu, "Regression Analysis of Longitudinal Data in the Presence of Informative Observation and Censoring Times," Journal of the American Statistical Association, vol. 102. pp. 1397–1406, 2007.

[7] S. R. Lipsitz, G. M. Fitzmaurice, J. G. Ibrahim, R. Gelber, and S. Lipshultz, "Parameter Estimation in Longitudinal Studies with Outcome-Dependent Follow-Up," Biometrics, vol. 58, no. 3, pp. 621–630, 2002.

[8] M. Sperrin, H. Rushton, W. G. Dixon, A. Normand, J. Villard, A. Chieh, and I. Buchan, "Who Self-Weighs and What Do They Gain From It? A Retrospective Comparison Between Smart Scale Users and the General Population in England," J. Med. Internet Res., vol. 18, no. 1, p. e17, Jan. 2016.