The University of Auckland
Examiner's Report on Masters Thesis / Research Portfolio

Candidate Name: Junhuang Xue **Student ID: 6365553** **Degree: STAT-MSC**
Thesis title: Anomaly Detection in Patient Arrival, with Bayesian Hierarchical Models

Part 2 – Comments for student

This section will be released to the candidate at the completion of the examination process

The examiner is requested to provide feedback for the student and also to list any minor editorial changes that they consider should be made before the thesis is lodged in the Library. Please note that in order to accommodate your detailed and comprehensive comments, the boxes below are expandable.

Summary of the contribution made by the thesis/research portfolio and the key strengths and weaknesses of the work

Some specific comments are:

This thesis begins by talking about (hierarchical Bayes models) HBM and the difficulties of too coarse or too fine ICD categorisations. Next, it launches into congestion at hospitals and anomaly detection. At this point it is difficult to perceive the link between these two endeavours and, indeed, a link is not well made.
The author then returns to HBM with a discussion of Bayes models.

In 1.6 the idea of hierarchical time series in introduced. The variable "y" is used to represent *something*, but it is unclear what. Examples are needed otherwise there is no context. Presumably as y is later modelled as Poisson, it the count of something occurring in an interval of time t. A count of what? And what is the time interval of interest?

2. Simulations. 2.2 describes hierarchy simulations of two categories A and B. What A and B might actually be in the real world is left unsaid. Again, it cries out for context and examples.
Next 2.3 a Poisson model for y (whatever y might be) is described and seems to have nothing much to do with the hierarchy simulations in 2.2. The categories A and B have entirely disappeared from the theory outlined in 2.3. Further, a parameter ρ is introduced which seems crucial to the development, but what it is is entirely unclear. E.g in "Model 1" (foot of page 27) it seems as if ρ and μ are the same thing. It is important to explain what ρ is.

Plots of prior distributions are shown in Fig 2.6. Presumably – though it is not said – these are prior for the λs?
In 2.3.2 there is an "important note" about the added anomaly. This is quite unclear. It should firstly not be buried here (given that "anomaly" is in the thesis title). Besides I can see nothing in the 2.3.3 Results that says anything about the anomaly, whether the anomaly is "detected" or not.
However I find in 2.4 the same "important note" and that 2.4 is about anomalies. It should be made clear in mathematical terms what the added anomaly (% of total) actually is and how it is represented. The means in Table 2.6 and 7 go up directly proportional to % anomaly. Why? Is this due to parameter k. eg. Anomaly25 is k=0.8= 1/1.25? But what "Heat maps" actually represent is a mystery to me.

In 2.3.3 posterior distributions of the models are presented. Again, there is no indication of what these distributions (Figs 2.6 and 2.7) actually represent. Comparing with the priors in Fig 2.5, I am confused by the different scales of measurement. The posteriors have a range ~5 - 25 (for whatever it is that is being represented) while in Fig 2.5 the range is 0-5. Why does the data dominate to throw the posterior so rightward?

The University of Auckland
Examiner's Report on Masters Thesis / Research Portfolio

Candidate Name: Junhuang Xue	**Student ID: 6365553**	**Degree: STAT-MSC**
Thesis title: Anomaly Detection in Patient Arrival, with Bayesian Hierarchical Models		

Table 2.1 is pairwise comparison of model DIC difference. These are very difficult to make sense of. I do not know why absolute values of DIC are not presented.

Chapter 3 is about real data. It concerns ICD codes and I finally get to know that y is (I think) about daily hospital admissions. I think the author is suggesting models for particular ICD code (at least I think this is what subscript I is?) .

On page 78 3 scenarios are identified and "anomaly detection rate" mentioned. This "rate" seems to me to be a crucial measure of the methodology, yet in 3.9 Results I can see no reporting of it. There are table concerning posterior distributions, but it is hard to see how these pertain to "rate" of detection.

Expand box as required

Criteria	Examiner's comment
Show evidence of advanced knowledge about a specialist field of enquiry	Possibly
Demonstrate mastery of sophisticated theoretical subject matter	Not really
Evaluate critically the findings and discussion in the literature	No
Research, analyse and argue from evidence	Lacks clarity
Engage in rigorous intellectual analysis, criticism and problem-solving	Not demonstrated
Demonstrate a high order of skill in the planning, execution and completion of an original, independent research project	No
Apply research skills appropriately	Not demonstrated

Recommended minor editorial corrections to thesis, if any
(not applicable to research portfolios)

There are typos to consider fixing. Some I noticed
P12. O.051 and 0.51 , which is it?
Formula 1.11. pD or Pd? Which is it?
The word "brunch" and "brunching" is used. I assume "branch" is intended.