Report

Signal detection of spontaneous medical device reports over time accounting for multiple comparisons

Ty Stanford, Lan Kelly, et al.

Table of contents

https://academic.oup.com/jamia/pages/General_Instructions

Research and Applications

Research and Applications articles describe original work in the formulation, implementation, or evaluation of informatics-based studies and investigations. The articles do not need to be limited to hypothesis-driven research, and they can, for example, report on an innovative application of information technology, the detailed description of a new methodology, or the formulation and formative evaluation of a new model. The structured abstract should contain the headings: Objective, Materials and Methods, Results, Discussion, and Conclusion. The main text should, in addition to the sections corresponding to these headings, include a section describing Background and Significance.

(Word count: up to 4000 words. Structured abstract: up to 250 words. Tables: up to 4. Figures: up to 6. References: unlimited.)

Brief Communications

Brief Communications are short versions of Research and Applications articles, often describing focused approaches to solve a particular problem, or preliminary evaluation of a novel system or methodology.

Word count: up to 2000 words. Structured abstract: up to 150 words. Tables: up to 2. Figures: up to 3. References: unlimited.

1 Abstract

Objective

Materials and Methods

Results

Discussion

Conclusion

2 Introduction

Adverse events from implantable medical devices are commonly reported to regulatory bodies in the form of unstructured free-text in spontaneous reports. Detecting safety signals from the reports for post-market surveillance can be challenging. Spontaneous reports of adverse events may be submitted by manufacturers, clinicians and consumers to the Database of Adverse Event Notifications (DAEN), maintained by the Australian Therapeutic Goods Administration (TGA) in the form of unstructured free text. Pelvic (urogynaecological) mesh was used to treat women for pelvic organ prolapse and stress urinary incontinence. While the device provided benefits to some women, others experienced serious complications.

The pelvic mesh was removed from market in …

There is still ongoing heath, financial and legal fallout from the device's use Dec 22

[talk about standard signal detection in structure databases]

[has there been any NLP of free-text in adverse events/safety monitoring?]

The purpose of this study was to develop an end-to-end pipeline, utilising free-text information to screen for medical device safety issues, using an Australian spontaneous report database of medical device adverse events (Database of Adverse Event Notifications, DAEN). The study aims to evaluate the feasibility of disproportional reporting rate methods to detect signals in free-text descriptions of adverse events using natural language processing (NLP) to classify DAEN reports into topics by:

1. implementing over time, repeated look adjusted disproportional reporting rate ("signal detection") algorithms on medical device adverse event data,
2. reporting retrospective time-to-signal findings for pelvic mesh devices associated with pain adverse events for analysis pipeline feasibility,
3. evaluating the sensitivity of signal detection methods to assumed data accumulation length and rates (e.g., the critical value calculation for maxSPRT), and
4. comparing the retrospective time-to-signal to the timings of the withdrawl of pelvic mesh devices in Australia.

3 Methods

The DAEN was searched for reports on pelvic and hernia mesh from 2012-2017. Topic modelling, a Natural Language Processing technique, was used to profile the unstructured text into mixtures of clinically relevant topics. A report was considered to contain a particular topic X if the probability it contains words relating to the topic, $\mathrm{P}(\text{topic} = X| \text{ document})$, was over a certain threshold, which was varied in the analysis.

Disproportionality analysis was used to detect potential signals from the most frequent clinical topic from pelvic mesh, with hernia mesh and other devices used as comparators. Measures are based on a 2×2 contingency table for the number of adverse events with and without the most frequent topic in the device of interest and the comparator. Testing was performed on the DAEN at quarterly intervals, if new data were accumulated in the interval, over the study period, commencing in 2012.

3.1 Data aquisition

The data is thanks to curtis-murray at his MedicalDevicesNLP repo

- Natural language processing of the TGA spontaneous reports of medical device database (DAEN)
- Each record has an estimate of P(`topic == "pain"` | `Level`, `Doc`) using hierarchical stochastic block modelling (hSBM)
- P(`topic == "pain"` | `Level`, `Doc`) estimates for each record are roughly interpreted as the proportion of the NLP analysed free text that is considered as using/describing words related to pain

And example record and processing values (description limited to 150 characters):

ReportID	Report date	Class	Device	P('pain'|doc)	ARTG no.	Event	Source	Event type	Description
37537	2015-02-06	other_device	Class IIa	0.029	137859	Injury	Industry	Mechanical	Patient admitted for routine SFA angioplasty. The physician had completed the procedure without incident and had withdrawn the balloon catheter fro…
36797	2015-08-27	other_mesh	Class III	0.020	219240	Injury	Industry	Material	3 weeks post-op, patient contacted the surgeon saying that the wound was opening, and she could see the implant. The patient sent photos, and there…

ReportID	Report date	Class	Device	P(‘pain’|doc)	ARTG no.	Event	Source	Event type	Description
40917	2016-04-24	pelvic_mesh	Class IIb	0.538	92718	Injury	Consumer	Other	Have had pelvic pain, pain with sex incontinence with bowel and bladder, cannot sit, walk, stand on feet for extended time.
45432	2017-03-29	hernia_mesh	Class IIb	0.214	98833	Injury	Consumer	Other	Atrium mesh implanted in my abdomen. Severe right sided abdominal pain ongoing. Also had further surgery to repair hernia due to reoccurrence of he…

	Report				ARTG			Event	
ReportID	date	Class	Device	P(‘pain’|doc)	no.	Event	Source	type	Description
44402	2017-12-01	other_devices	Class III	0.000	149128	No Injury	Other	Other	Sponsor distributed a Customer Letter dated 20th December 2016 announcing the immediate discontinuation of their CE-marked Umbilical Vessel Cathete...
45624	2017-12-04	other_devices	Class 1	0.000	121950	No Injury	Health Professional	Mechanical	Operating table started tilting patient on its own during procedure. Emergency stop button pressed - table stopped briefly and began to tilt patien...

ReportID	Report date	Class	Device	P(‘pain’|doc)	ARTG no.	Event	Source	Event type	Description
45265	2017-12-18	other_device	Class IIb	0.250	177101	Injury	Industry	Other	Pain, possible rupture and swelling around prosthesis.

3.2 Analysis data

Signal detection of disproportionate adverse events (AEs) will often have tabulated count data accumulated over time. The data at time point t can be summarised as below:

	AE(s) $\in Y$	AE(s) $\in \bar{Y}$
Target exposure	a_t	b_t
Comparator exposure	c_t	d_t

where

- AE(s) Y is the set of AEs (or singular AE) of interest,
- AE(s) $\bar{Y}$ is the complementary set to the AEs of interest,
- *Target exposure* is the medical device(s) of interest,

- *Comparator exposure* is the medical devices to which the *Target exposure* is being compared, and
- a_t, b_t, c_t and d_t (all $\in \mathbb{Z}^+$) are the respective counts of AEs recorded up until (i.e., cumulative) time t.

In the motivating example of the pelvic mesh device, the contingency table can be written more specifically as

	Pain AEs	Not pain AEs
Pelvic mesh	a_t	b_t
Comparator exposure	c_t	d_t

where

- *AEs pain* is the count of AEs that contain "pain" themes greater or equal to some pre-specified threshold $p_t \in (0, 1)$ as estimated by the hSBM (that is, $P(\texttt{topic == "pain"}|\texttt{Level, Doc}) \geq p_t$), and
- *Comparator exposure* can be any relevant set of medical devices to compare the pelvic mesh to (e.g., hernia mesh or all other mesh devices or all other devices).

3.3 Signal detection over time

We will consider the three signal detection statistics below:

- Proportional reporting ratio (PRR),

- Bayesian Confidence Propagation Neural Network Information Component (BCPNN IC with MCMC CIs), and
- the maxSPRT statistic

As signal detection is being undertaken repeatedly as data are being accumulated, alpha spending needs to be considered. The below table classifies the aforementioned signal detection methods by their null hypothesis as well as whether they control for the family-wise error rate (FWER)

Null hypothesis	non-FWER version	FWER version
Ratio of pain AEs to all AEs in target and comparator groups has a ratio of 1	PRR	binary, group sequential maxSPRT
Independence of pain AEs and target group (based on marginal counts)	IC	IC with α-spending scheme

We will demonstrate how the group sequential binary maxSPRT, as described in previous work, is equivalent to a FWER-controlled PRR method of signal detection.

Methods used included the Bayesian Confidence Propagation Neural Network (BCPNN) and the maximised Sequential Probability Ratio Test (maxSPRT) which accounted for multiple testing through alpha spending. The BCPNN was used with and without adjusting for multiple testing. The test statistic for BCPNN is the information criterion (IC) which represents the log2 of the ratio of observed to expected adverse events (0 under the null hypothesis of no association between the topic and pelvic mesh).

maxSPRT was developed for near continuous sequential monitoring (called "group sequential" when monitored at discrete time points or after set accumulatios of events), maintaining the correct overall alpha level. The test statistic is based on the maximized (log-)likelihood ratio statistic which uses the observed and expected (under the null hypothesis) reporting ratio assuming binomial adverse event accumulation. The critical value of the likelihood ratio statistic is determined by the 100(1-)% quantile of possible likelihood ratio statistics from binomial adverse event counts under the null hypothesis over the entire group sequential follow-up.

3.3.1 Proportional reporting ratio (PRR)

The PRR estimate is calculated

$$\widehat{\mathrm{PRR}}_t = \frac{\frac{a_t}{a_t+b_t}}{\frac{c_t}{c_t+d_t}}.$$

In the context of signal detection, an elevated proportional reporting ratio is of concern. Therefore the one-sided hypothesis test $H_0 : \mathrm{PRR} \leq 1$ (proportional reporting of the target is less than the comparator) is used and is not rejected until

$$\widehat{\mathrm{PRR}}_t \times \exp\left\{-Z^*_\alpha \sqrt{\frac{1}{a_t} + \frac{1}{a_t+b_t} + \frac{1}{c_t} + \frac{1}{c_t+d_t}}\right\} > 1$$

at the α level where Z^*_α is the $(1-\alpha)^{\mathrm{th}}$ quantile of the standard normal distribution. The above threshold is equivalent to the lower bound of the approximate $100(1-2\alpha)\%$ confidence interval for a standard two-sided hypothesis test.

3.3.2 Bayesian Confidence Propagation Neural Network (BCPNN) Information Component (IC)

The Information Component (IC) statistic is an estimate of the observed-to-expected ratio of the number of target exposure AEs of interest on the $\log_2$-scale under independence between the target exposure and AEs of interest based on information theory (Bate et al., 1998)

$$\mathrm{IC}_{XY} = \log_2 \frac{P_{X,Y}(a_t+b_t, a_t+c_t)}{P_X(a_t+b_t)P_Y(a_t+c_t)}$$

where $P_X(X = x)$ denotes the marginal probability of an observed count x for the target exposure, $P_Y(Y = y)$ denotes the marginal probability of an observed count y for the AE of interest, and $P_{X,Y}(X = x, Y = y)$ denotes the joint probability.

The BCPNN IC of Noren et al. (2006) uses a Bayesian inference based *maximum a posteriori* (m.a.p.) central estimate of the IC,

$$\widehat{\mathrm{IC}}_t = \log_2 \frac{\mathrm{E}\left[\hat{p}_a\right]}{\mathrm{E}\left[\hat{p}_a + \hat{p}_b\right]\mathrm{E}\left[\hat{p}_a + \hat{p}_c\right]}$$

where p_a, p_b and p_c are the (assumed constant over time) underlying probabilities of the multinomial-distributed observed events a_t, b_t and c_t, respectively (p_d corresponding to the count d_t also included). The underlying probabilities are modelled using Dirichlet priors resulting in a Dirichlet posterior distribution. The one-sided null hypothesis of the joint probability target exposure and AEs of interest is equal or less than the marginal products ($H_0 : \mathrm{IC}_t \leq 0$) can be rejected when the α quantile of the Markov Chain Monte Carlo (MCMC) empirical distribution is greater than 0. Similarly to the rejection rule for the PRR, this threshold corresponds to the lower bound of the $100(1-2\alpha)\%$ equal-tailed credible region in a two-sided hypothesis test.

3.3.3 maxSPRT

Kulldorff et al. (2011) outlined that the relative risk (RR) at a given point-in-time for accumulated binary data (that is, “success”/“failure” events or AE of interest or not) of a target group relative to a comparator has the maximum likelihood estimate of

$$\widehat{\mathrm{RR}} = z\frac{C_n}{n - C_n}$$

where

- z is the ratio of the total AEs for the comparator to the total AEs for the target,
- C_n is the count of target exposure AEs in X,
- n is the count of all AEs in X (target and comparator exposure), and
- $n - C_n$ is therefore the count of comparator exposure AEs in X.

In the context of our data, the values z, C_n and n are the quantities $\frac{c_t+d_t}{a_t+b_t}$, a_t and $a_t + c_t$, respectively, at time t.

Therefore the RR maximum likelihood estimate at time t can be re-written

$$\begin{aligned}\widehat{\mathrm{RR}}_t &= \frac{c_t + d_t}{a_t + b_t} \times \frac{a_t}{c_t} \\ &= \frac{\frac{1}{a_t+b_t}}{\frac{1}{c_t+d_t}} \times \frac{a_t}{c_t} \\ &= \frac{\frac{a_t}{a_t+b_t}}{\frac{c_t}{c_t+d_t}}\end{aligned}$$

which is the PRR estimate at time t as before.

The (maximised) log-likelihood ratio statistic of $\widehat{\mathrm{PRR}}_t$ (equivalently, $\widehat{\mathrm{RR}}_t$) can be determined calculated as

$$\mathrm{LLR}_t = a_t \ln\left(\frac{a_t}{a_t + c_t}\right) + c_t \ln\left(\frac{c_t}{a_t + c_t}\right) - a_t \ln\left(\frac{a_t + b_t}{a_t + b_t + c_t + d_t}\right) - c_t \ln\left(\frac{c_t + d_t}{a_t + b_t + c_t + d_t}\right)$$

The maxSPRT test is considered significant when LLR_t is greater than the pre-computed critical value which is the $100(1-\alpha)\%$ percentile of the LLR_t values generated under the null hypothesis $\mathrm{RR}_t = 1$ for group sequential looks at the data $t = t_1, t_2, \ldots, t_k$. The CV can either be computed using the 95th percentile of LLR_t values with the exact joint binomial probabilities over k looks of the data accumulation under the null hypothesis, or by MCMC sampling of binomial event accumulation (and associated LLR_t values) to approximate the LLR distribution when the exact CV computation is computationally intractable.[1]

[1] The `Sequential` R package exact CV function suggests ”not using values greater than 1000” in regards to

3.4 Analysis choices

A large unknown what threshold should be used to dichotomise P(`topic == "pain" | Doc`) into pain and non-pain spontaneous event report. The threshold can roughly be interpreted as the proportion of the spontaneous event report free text relates to "pain" topics. It is a balancing act, likely with some "safe zone", to choose a threshold high enough that false-positive pain reports don't occur too frequently to induce noise or bias that might exist between the two device groups being compared, and also importantly a threshold low enough that pain events are not missed with false negatives and having the flow on effect of not having enough events of interest to sufficiently power the disproportionality statistics. From pilot feasibility exploration, the thresholds considered were from 0.5% to 10% of the free text, i.e., using thresholds $e_{topic} = \{0.005, 0.010, ..., 0.100\}$. Further complicating matters, it should be noted, optimal P(`topic == "pain" | Doc`) thresholds may be different for differing event topics, however, we only consider the pain topic in this study.

We choose to perform the analysis group sequential analysis at quarterly time points as this is a reasonable accumulation of events in the Australian specific data source and represents a data accumulation interval feasible for large scale medical device safety monitoring given the computational and pipeline complexity of NLP of the free text and downstream analysis.

There are a vast amount of options in choosing an alpha-spending function to use with the BCPNN IC (lower) confidence interval calculations. However, we have chosen the widely used exponential spending function (Anderson and Clark, 2009) with $\nu = \frac{1}{2}$ because of its wide use and reasonable properties.

The statistical analysis enumerates the following choices:

- comparator: pelvic mesh is compared to a range of specific (hernia mesh) to less specific (all other devices) comparator groups to characterise analysis performance,
- pain topic threshold, and
- competing signal detection methods in alpha-spend adjusted BCPNN IC calculations and maxSPRT.

[2]

the total events to potentially be observed over the follow-up.

[2] Anderson KM and Clark JB (2009), Fitting spending functions. Statistics in Medicine; 29:321-327. Jennison C and Turnbull BW (2000), Group Sequential Methods with Applications to Clinical Trials. Boca Raton: Chapman and Hall. Lan, KKG and DeMets, DL (1983), Discrete sequential boundaries for clinical trials. Biometrika; 70:659-663.

4 Results

Pelvic mesh had the highest proportion of reports including the word “pain”, followed by other and hernia mesh (Figure 1).

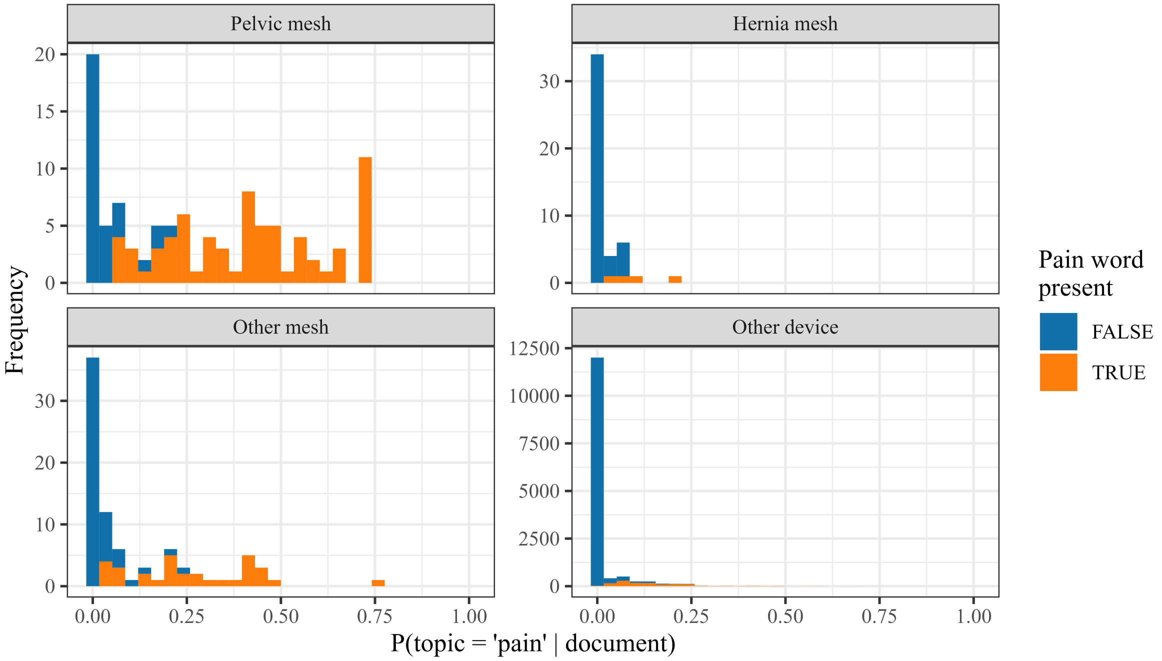

Figure 1: Frequency of pain topic in pelvic, hernia and other mesh and other devices.

The cumulative number of pain topics in pelvic mesh a_t increased sharply in Q4 2014 compared with the previous quarter. The table shows the cumulative 2×2 table each quarter up to $t =$ 2015-Q1 with hernia mesh as the comparator (and pain topic threshold of 0.05).

Table 5: Cumulative quarterly AE counts of pelvic mesh (a_t is pain topic count) compared to hernia mesh (c_t is pain topic count) using a pain topic threshold of 0.05.

Quarter	t	a_t	b_t	c_t	d_t
2013-Q3	1	4	12	1	10
2013-Q4	2	6	14	1	10
2014-Q1	3	6	14	1	11
2014-Q2	4	7	15	1	14
2014-Q3	5	9	17	3	21
2014-Q4	6	26	19	4	27
2015-Q1	7	27	19	4	28
2015-Q2	8	27	19	4	28

Quarter	t	a_t	b_t	c_t	d_t
2015-Q3	9	27	20	4	28
2015-Q4	10	27	20	6	28
2016-Q1	11	30	21	6	28
2016-Q2	12	34	21	6	28
2016-Q3	13	34	21	7	33
2016-Q4	14	36	23	7	33
2017-Q1	15	45	23	8	34
2017-Q2	16	58	24	8	37
2017-Q3	17	68	24	8	38
2017-Q4	18	77	25	8	38

Figure 2 demonstrates …

While not the values used for determining significance, they are the statistics that are most familiar to interested parties. The RR estimates for maxSPRT and PRR methods are of course identical, the methods differ in using the LLR and Wald testing respectively. The IC statistic is not directly comparable to RR, it does share similarities in that values above the null value represent disproportionate reporting rate in the cohort of interest. (include a 2^IC vs RR numeric example here)

The columns of the plot depict increasing pain topic thresholds. Within each threshold, the reporting ratio estimates are reasonably stable after some initial fluctuation as a result of the uncertainty discrete nature of the data accumulated over time. However, the RR estimates for the larger thresholds show a general increase in the reporting ratio estimates over time. Interestingly, this is not the case for the BCPNN which actually decreases for the 0.08 threshold from 2016 onward. Common to all methods though are increased reporting ratio estimates for larger thresholds within their respective method suggesting that larger P(topic = “pain”| document) are more frequent in the pelvic mesh group. This higher threshold comes with a cost, that the higher barrier to pain events means insufficient data are available to produce a reporting ratio estimate early in the data accumulation. Further demonstrating competing constraints of a threshold sensitive enough to identify pain events and …

Note these stats are not affected by multiple comparison adjustment

Vertical lines are when H0 is rejected for the various methods which don’t necessarily coincide with the largest reporting ratio estimate but with the …

Significance was reached at the same time for all methods (last quarter of 2014), except for the PRR method at the 0.08 threshold which was a delayed an additional quarter.

Figure 3 shows how significance is determined for the three methods. The vertical line is when significance is reached

LLR valuyes have been truncated at 30 for

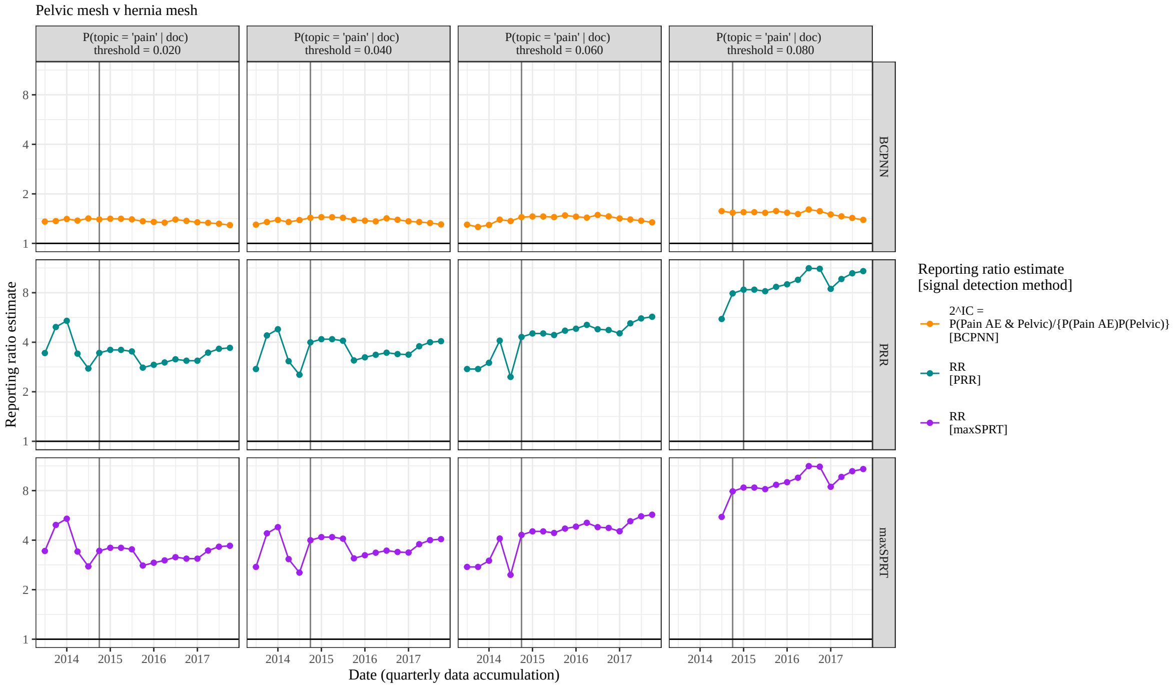

Figure 2: Reporting ratio estimates for the pelvic mesh v hernia mesh comparison over time for the three signal detection methods at different pain topic thresholds.

The CV values for each maxSPRT analysis for the different comparators are different (values here), although not obvious from the graph

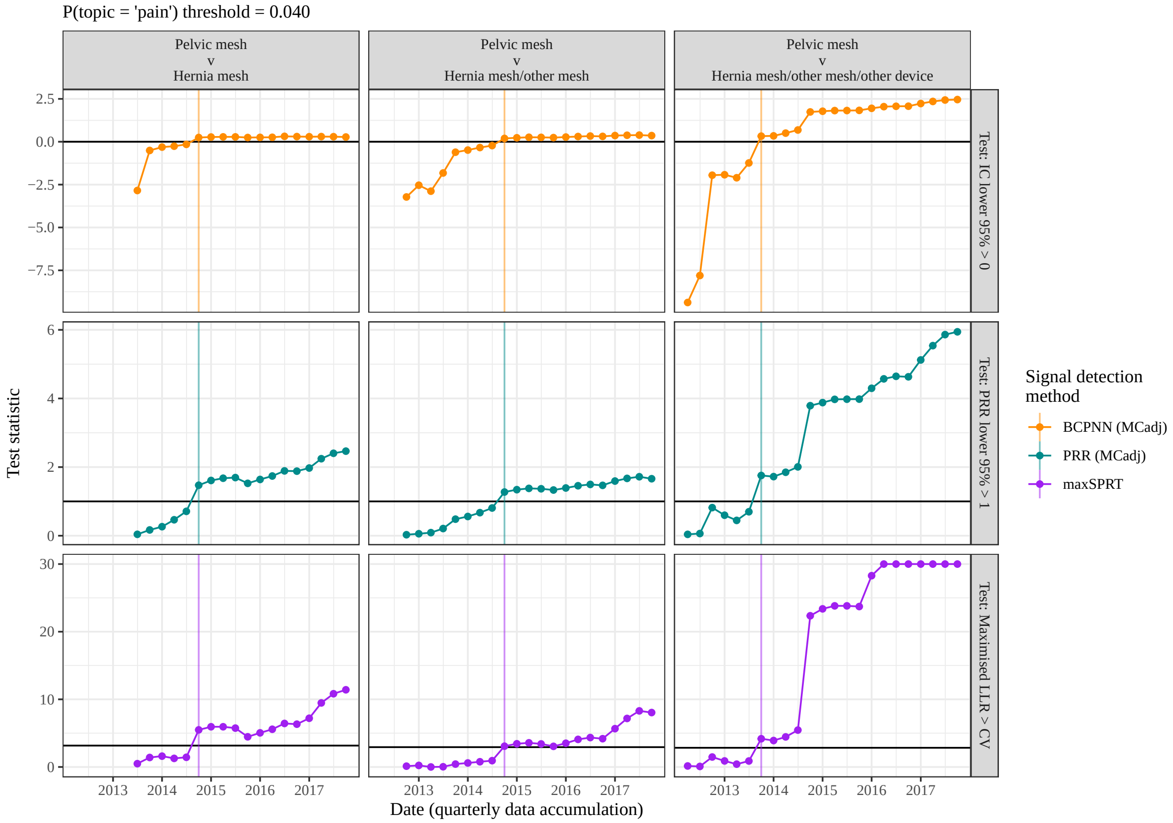

Figure 3: Disproportionality test statistic values for heneria mesh compared to three copmparison groups over time with their respective critical values. The P(topic = 'pain'| document) threshold is set at 0.04.

Figure 4 shows how the different signal detection methods reached significance at different pain thresholds.

The highest thresholds risk delayed signal detection or not reaching significance at all, especially in the least data rich comparison of pelvic mesh v hernia mesh.

xxx where the thresholds that ensured all methods for the mesh only comparisons were detected prior to 2015

while the prr allowed earlier detection in the pelvic v all comparison (2012 Q4) for thresholds xxxx, these thresholds were sub-optimal for the other methods of signal detection

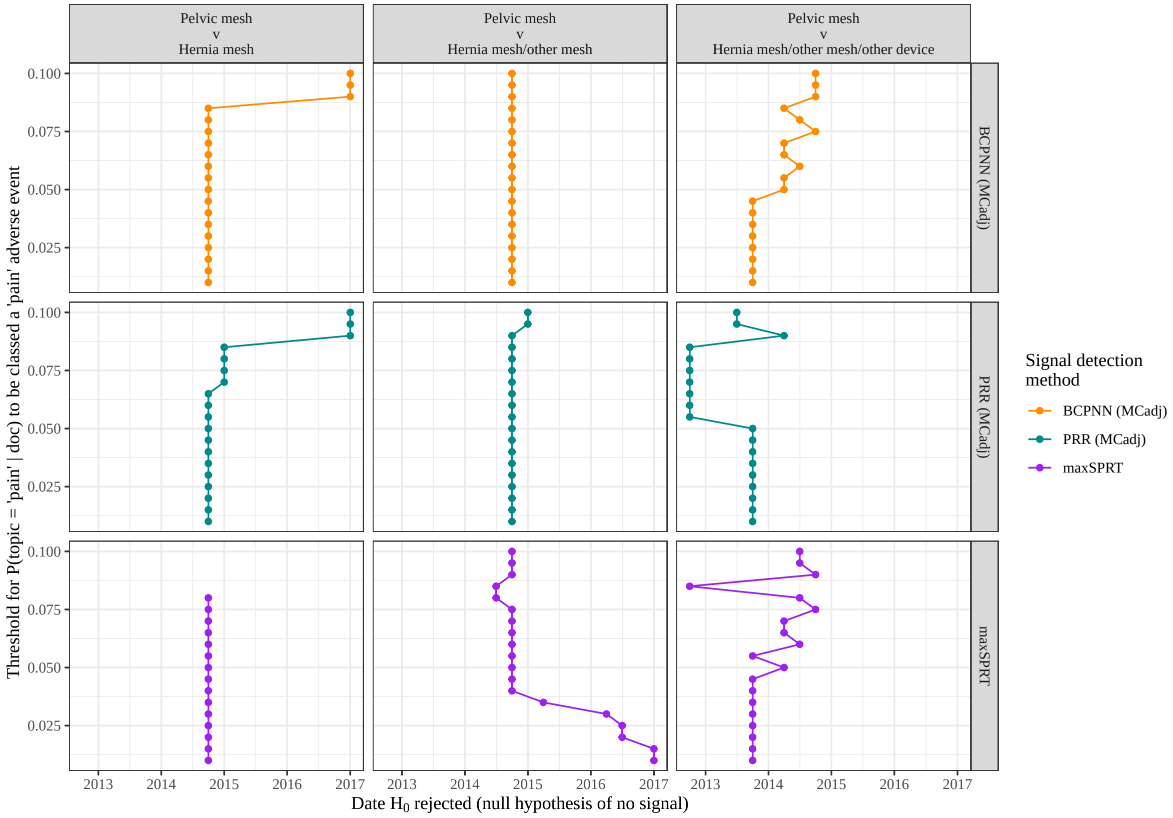

Figure 4: Time when methods reached critical values for signal detection.

5 Conclusion

Urogynaecological mesh was withdrawn from the Australian market in 2018, while our retrospective analysis with a 0.06 threshold detected signals between August – December 2014. We have demonstrated the potential of using topic modelling in spontaneous reports for signal detection in post-market surveillance.

6 Session information

```r
format(Sys.time(), '%d %b %Y')
```

```
[1] "04 Dec 2023"
```

```r
sessionInfo()
```

```
R version 4.3.1 (2023-06-16 ucrt)
Platform: x86_64-w64-mingw32/x64 (64-bit)
Running under: Windows 10 x64 (build 19045)

Matrix products: default


locale:
[1] LC_COLLATE=English_Australia.utf8  LC_CTYPE=English_Australia.utf8
[3] LC_MONETARY=English_Australia.utf8 LC_NUMERIC=C
[5] LC_TIME=English_Australia.utf8

time zone: Australia/Adelaide
tzcode source: internal

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
 [1] arrow_12.0.1.1  gsDesign_3.5.0  knitr_1.43      ggrepel_0.9.3
 [5] ggthemes_5.0.0  ggplot2_3.4.2   stringr_1.5.0   lubridate_1.9.2
 [9] forcats_1.0.0   tidyr_1.3.0     dplyr_1.1.2     readr_2.1.4

loaded via a namespace (and not attached):
 [1] gt_0.9.0          utf8_1.2.3        generics_0.1.3    xml2_1.3.5
 [5] stringi_1.7.12    hms_1.1.3         digest_0.6.33     magrittr_2.0.3
 [9] evaluate_0.21     grid_4.3.1        timechange_0.2.0  fastmap_1.1.1
[13] jsonlite_1.8.7    purrr_1.0.1       fansi_1.0.4       scales_1.2.1
[17] textshaping_0.3.6 cli_3.6.1         rlang_1.1.1       bit64_4.0.5
[21] munsell_0.5.0     withr_2.5.0       yaml_2.3.7        tools_4.3.1
[25] tzdb_0.4.0        colorspace_2.1-0  assertthat_0.2.1  vctrs_0.6.3
[29] R6_2.5.1          lifecycle_1.0.3   bit_4.0.5         ragg_1.2.5
```

```
[33] pkgconfig_2.0.3   pillar_1.9.0      gtable_0.3.3      glue_1.6.2
[37] Rcpp_1.0.11       systemfonts_1.0.4 xfun_0.39         tibble_3.2.1
[41] tidyselect_1.2.0  rstudioapi_0.15.0 farver_2.1.1      xtable_1.8-4
[45] htmltools_0.5.5   labeling_0.4.2    rmarkdown_2.23    compiler_4.3.1
```