Research Publication

The Case Quantify and Search Tool (C-QST)

An automated tool to transform case notes into quantitative information

Chee Seng Chong, Alessandra Raudino, Ofir Thaler & Mark Howard

Research Publication No. 56
December 2017

ISSN 0813 5800

Corrective Services NSW

Key Summary

Background. A systematic examination of the qualitative data embedded in case notes can provide insight into the dynamics of offender case management and intervention. However, the size of these datasets has far exceeded the capacity for standard data processing and qualitative analysis methods. Therefore, the development of new procedures and research tools that incorporate 'Big Data' processing techniques, such as data and text mining methods, is required.

Aim. This paper documents the development of the Case Quantify and Search Tool (C-QST), an instrument that uses text mining and natural language processing techniques to automatically convert case note content into quantitative data. It discusses the C-QST's program logic, validation, potential uses and implications for research.

Method. A sample of 84,115 Practice Guide for Intervention related case notes generated by Community Corrections Officers between June 2016 and September 2017 were extracted from the Corrections Services New South Wales (CSNSW) Offender Integrated Management System. Pre-processing of the data resulted in a final data set of 82,500 case notes.

Findings. The validity of the C-QST was first established on a sub-sample of 50 case notes. Cohen's Kappa revealed an optimal agreement between the automated results and a manual review conducted by the authors of this paper. The C-QST was able to identify and quantify qualitative information embedded within the content of case notes, providing the kind of detailed insights that are normally only accessible through a manual qualitative review.

Conclusion. Developed by CSNSW, the C-QST is the first automated search tool of its kind. Results from its initial application seem to corroborate the utility of novel approaches for data triangulation and data mining techniques in CSNSW.

Contents

List of Tables

List of Figures

Introduction

During the last decade, there has been increasing interest in the use of textual and qualitative information stored in large administrative and routinely collected datasets and their triangulation with quantitative indicators. Triangulation commonly refers to the use of multiple data sources with the aim of extracting adequate meaning from the data and enhancing the quality of inferences made. However, the complex and unstructured nature of these datasets requires a sophisticated methodology and database architecture to process and maximise their systematic use.

Recently, technical advances in software innovation and computer programming language have encouraged a more innovative approach to developing and reconciling qualitative and quantitative methods and triangulating different data sources. The relative cost effectiveness of gathering and maintaining a large repository of data has also significantly encouraged rapid growth in the size and inclusiveness of datasets. These 'Big Data' have far exceeded the capacities for standard manual analysis procedures and have, in part, cultivated the current research climate that places emphasis and value in the understanding of large data through innovative techniques such as data mining and machine learning methods. These methods are designed to extract information from large volumes of data through the development of algorithms, decision trees or cluster analyses to disclose the hidden pattern within unstructured data. Data mining allows the analysis of large and complex information that is not suitable for hand database management tools or traditional data processing applications.

To address the need to utilise a large amount of qualitative information embedded in routinely collected administrative databases, this paper presents an innovative approach that utilises data and text mining techniques for converting free text contained in case notes into quantifiable information for further analysis.

The Present Study

In Corrective Services New South Wales (CSNSW), case notes are a complete record of an offender's contact with the corrective system that are written and stored as text-based resources on a comprehensive operational database known as the Offender Integrated Management System (OIMS). Case notes help support supervision and decision-making by

ensuring that staff coming into contact with an offender, or reviewing a case, has access to relevant information.

Case notes are traditionally seen as administrative tools for record keeping purposes; however, the analysis of case notes may provide invaluable insights into what is often a 'black box' of offender case management and supervision. A challenge associated with such analysis is that not only is it difficult to extract meaningful information from unstructured data, but also the number of case notes have long exceeded the capacity for standard data processing and qualitative analysis methods. In the local context, an additional consideration is that the size of the data are increasing at an accelerated rate due to growth in the populations of offenders supervised in custody and in the community (e.g., Raudino, Neto & Van Doorn, 2017).

It is within this context that the Case Quantify and Search Tool (C-QST) was created. The C-QST utilises text mining and natural language processing techniques to automate the process of converting case notes into quantifiable data. The strength of the tool lies in how unstructured open case note texts can be restructured into a meaningful order, allowing for the application of rules to heuristically identify relevant information. For the purposes of this report, the utility of the C-QST is demonstrated through its application to case notes generated as part of the routine implementation of the Practice Guide for Intervention (PGI).

The PGI comprises a series of exercises that that are undertaken with community-based offenders as part of their supervision by CSNSW Community Corrections. Designed as a structured intervention that assists Community Corrections Officers in applying motivational interviewing and cognitive behavioural techniques to address offenders' criminogenic needs, an exercise may contain up to three worksheets. Exercises that address the same criminogenic need are grouped into modules. In total, the PGI consists of 13 modules, 56 exercises and 78 worksheets (see CSNSW, 2016).

The delivery of PGI exercises in supervision sessions are recorded on OIMS; specifically, the content of the supervision and the PGI module are noted. The content of the supervision is entered in an open text format while the PGI module is indicated via the selection of the appropriate value from a dropdown list (e.g., 'Assessment and Planning', which corresponds to PGI Module 1). Thus, while PGI module information is systematically documented on OIMS, more detailed information about implementation, such as the PGI exercises used, are

not quantitatively recorded through any coding or categorisation system on OIMS. This level of information is embedded within the content of case notes and can only be gained through a manual qualitative review, which is both subjective and time-consuming.

This paper aims to describe the development of the C-QST and demonstrate its utility at identifying and quantifying latent information embedded in the content of case notes, such as PGI exercise use, providing insights at a finer level of detail than otherwise possible. In this demonstration, all case notes relating to the delivery of PGI sessions from 2 June 2016 to 5 September 2017 were extracted from OIMS and processed by the tool. The aims of the C-QST in this demonstration were to:

- cross validate free text case note content (output from the C-QST) against module category information recorded on OIMS
- examine implementation of the PGI at the exercise level.

Method

Sample

A total of 84,115 PGI related case notes were examined. From this sample, 1,164 case notes were removed, as they were identified to be non-PGI related (i.e., PGI-Other). The category PGI-Other is used to document sessions that fall outside specific implementation of any of the 13 PGI modules, but during which behaviour-change conversations are conducted. A further 451 case notes were removed because they appeared to refer to offenders who were not supervised by Community Corrections. This resulted in a final data set of 82,500 case notes that were processed by the C-QST. The following section illustrates the C-QST program logic.

Procedure

The C-QST was developed in R (version 3.3.2). In its current form, it was designed to identify which PGI exercise and module was used in each case note. The first step required the user to import case note data extracted from OIMS directly into the R environment. All text characters were converted into lower case and all punctuations were replaced with a

white space character. The case notes were then converted into N-grams using the 'tidytext' package (Silge & Robinson, 2016). N-grams can be conceptualised as a method of data organisation in which the entire body of case note text is sliced into smaller units that consist of a contiguous sequence of N-items. Table 1 shows how a case note is converted into a 5-gram format. As can be seen, each row contains one 5-gram and all 5-grams are sequenced in descending order. Each sequence moves forward by one word, so there is always a 4-word overlap between rows.

Table 1. A Case Note Example in 5-Grams[1]

Word1	Word2	Word3	Word4	Word5
what	is	supervision	it	means
is	supervision	it	means	that
supervision	it	means	that	you
it	means	that	you	will
means	that	you	will	need

The use of an N-gram methodology offers several advantages. First, by decomposing a body of text into smaller parts, any unforeseen errors in the text will only affect a limited number of N-grams, leaving the rest of the text in the case note available for further processing. Second, N-grams provide a simple structure to order and store word-level information. As explained below, this allows for more efficient use of match and filter functions. Finally, N-gram methodology allows the search scope to be widened to words that precede and follow the selected search terms (e.g., the exercise names). These 'flanking' words provide contextual information in which rules can be applied to fine-tune the sensitivity of the search, balancing between false positive and negative rates.

Table 1 shows a sample of case notes divided into five columns, labelled Word1–Word5. Each column is parsed through to identify instances in which a word or number sequence in the case note matches the search terms. This search method is made up of four steps, as shown in Figure 1.

[1] The original case note read: 'What is supervision? It means that you will need …'

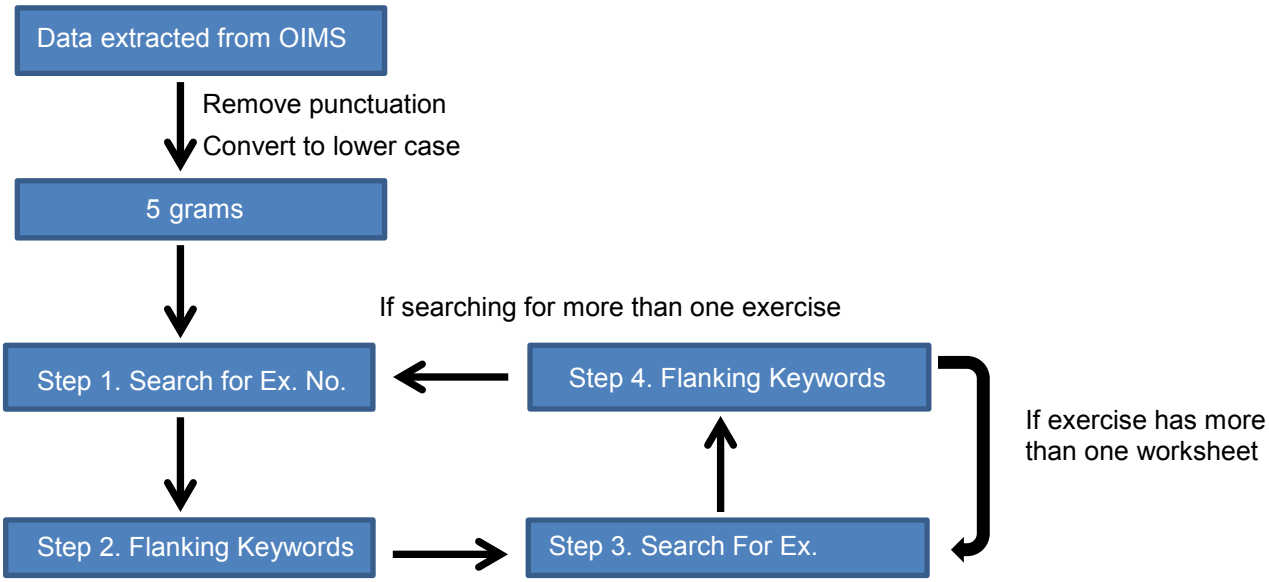

Figure 1. Flow Diagram of Program Logic

Each of these steps is discussed below. In this example, the tool only recognised the exercise as 'true' if specific words from Exercise 6.2 'Identifying High Risk Situations' were present in a meaningful sequence (see Table 2).

Table 2. Sample Case Note Exercise 6.2

Word1	Word2	Word3	Word4	Word5
on	6	2	2017	offender
6	2	2017	offender	completed
2	2017	offender	completed	worksheet
2017	offender	completed	worksheet	6
offender	completed	worksheet	6	2

Step 1. Search for Exercise Number Sequence

The aim of Step 1 is to identify whether the target exercise number sequence occurs in the case note. The targeted exercise number in this example (i.e., 6.2) was separated into two digits (i.e., 6 and 2). The data was then filtered to keep only instances in which the first digit (i.e., 6) appeared in column Word1 and the second digit (i.e., 2) appeared in Word2. In this case, only one row (i.e., the second row of Table 2) met these conditions, resulting in a reduced data set (see Table 3).

Table 3. Example Output from Step 1

Word1	Word2	Word3	Word4	Word5
6	2	2017	offender	completed

Step 2. Apply Flanking Keyword Rule

However, this search return is a false positive, as the numbers '6' and '2' refer to a date rather than a PGI exercise. The purpose of Step 2 is to determine the context of the search return by widening the scope of the search to examine the words that precede or follow the search term. A search return is only counted if one or more of the words that flank the search term matches a predefined keyword. Called 'flanking keywords', these are user defined and can easily be modified (see Table 4 for the list of flanking keywords used in the C-QST).

Table 4. Flanking Keywords

List of Flanking Keywords	
^complete	discuss
^ex	started
^mod	revisit
^undert	explain
^pg	assessment
sheet	canvass
conduct	^use
deliver	^give

Flanking keywords are treated as 'regular expressions' to account for common spelling mistakes, abbreviations, plurals and verb tenses (past and present tenses). For example, the expression '^mod' means that any word that starts with 'mod' will be recognised; in this case, not only will the word 'module' be accepted, but also its plural 'modules', abbreviation 'mod' and even misspelt variations such as 'modual' or 'modue'.

As shown in Table 3, since the word following the search terms ‘6’ and ‘2’ is ‘2017’—which is not on the flanking keywords list—this search return is not counted as a valid hit and is correctly rejected as a false positive.

Step 1 is then repeated with the filter applied to Word2 and Word3 (rather than Word1 and Word2), and Step 2 is repeated (if there are acceptable search returns). These steps are then repeated until either a valid hit is found or the search ends after parsing through columns ‘Word4’ and ‘Word5’.

In the event that a valid search hit is found, the case note is removed from further search iterations to prevent a case note from being counted twice. The tool then proceeds with the next two steps, searching for instances in which a case note may have been missed because an exercise name was reported in place of exercise number.

Step 3. Search for Exercise and Worksheet Names

Exercise and worksheet names are also treated as regular expressions (i.e., ‘regex’; the complete list is documented in the Appendix, see Table A1). As in Step 1, the sequence of words comprising an exercise name are separated and used to filter out instances in which a sequence matches the exercise name. This subset of case notes is then processed following Step 4, in which the flanking keyword rules are applied.

Step 4. Apply Flanking Keyword Rule

The decision to use 5-grams was based on the number of words the PGI exercise and worksheet names contain. This means that an N-gram with an N of at least 5 is required to apply the flanking keyword rules on PGI exercises with names that are made up of four words. Names that are longer than four words are truncated to four (e.g., Exercise 11.1 ‘What Are My Strengths and Skills?’ was truncated to ‘My Strengths and Skills’).

Steps 3 and 4 are repeated for each worksheet name associated with an exercise. These steps are repeated for each exercise that is entered as a search term. As an optional step, module numbers and module names can also be included to identify case notes that provide module only information.

Tool Validation

The validity of the tool was determined by examining the level of agreement between the tool and a manual review conducted by two independent raters. A subset of 15 case notes was randomly selected and reviewed by the authors of the paper to establish initial reliability. The independent raters were required to read through each case note to determine which PGI exercise was used. Complete agreement between the two raters was achieved and assessed through Cohen's Kappa values ($k = 1$, indicating perfect agreement). A total set of 50 case notes was then manually rated and compared to the output generated by the C-QST.

Cohen's Kappa values were used to compare the manual rating with the automated review results. Kappa values ranged from 0.73 to 1.00, indicating a good to optimal level of agreement between the raters and the tool. It is generally accepted that a value of Kappa from 0.60 to 0.79 indicates substantial agreement while 0.80 and above shows outstanding agreement (Landis & Koch, 1977).

Results

Cross Validation of C-QST Output against OIMS Module Information

Table 5 shows the distribution of case notes across the different modules. The column 'OIMS Modules' lists the 13 PGI modules; the adjacent column labelled 'Total No. Case Notes' shows the number of case notes that were generated for each module within the sample period. The remaining columns show the output from the C-QST.

The diagonal (in red) shows the proportion of case notes in which the tool identified evidence of exercise use that was represented in the module reported on OIMS. The column labelled 'Unknown' lists the proportion of case notes in which the C-QST was unable to identify any reference to PGI exercise use. For example, out of the 44,193 case notes that were recorded as Module 1 on OIMS, the C-QST was able to identify evidence of Module 1 exercise use in 84.1% of these case notes. However, the C-QST was unable to recognise any reference to PGI use in 15% of the case notes. These numbers (i.e., 100%, 84.1% and 15%) suggest that an estimated 0.9% of case notes contain evidence of exercise use from modules other than Module 1. These estimates are listed in the column labelled 'Mismatch'.

It was apparent that some case notes contained references to multiple PGI modules. If each case note only referenced exercises from a single PGI module, the estimated proportion of mismatched case notes for Module 1 (0.9%) should have been equal to the sum of all case notes that referenced exercises from other modules (2–13). This was clearly not the case; the C-QST found evidence of Module 2 use in 10.7% of Module 1 case notes. This indicates that there were instances in which exercises from both Module 1 and 2 were recorded using a single case note as a single module. As a consequence, the delivery of some PGI modules may remain hidden within case note content, which means that PGI module information on OIMS may be underestimating PGI use.

In general, the output obtained from the C-QST closely reflected the module information reported on OIMS. The performance of the C-QST appeared to be the least reliable for Module 12 'Pro-Social Lifestyle' (51%), which also had the largest proportion of case notes in which the C-QST was unable to detect evidence of PGI use (38.9% Unknowns).

On average, 23.5% of case notes were identified as not including explicit reference to any exercise (i.e., Unknowns) while an average of 4.8% were identified as having mismatched modules. Some of these discrepancies may be due, in part, to the limitations of the tool (discussed further below);[2] however, further exploration of the data (see Table 6) shows that the proportion of Unknowns appeared to decrease as a function of time. This suggests that some Unknowns and Mismatched case notes may be a result of transient reporting issues associated with the initial rollout of the PGI.

[2] This analysis was conducted prior to the release of the *Guide to Good Case Notes* (Corrective Services New South Wales, 2017), which stipulates the need to include in case notes the name of the exercises and worksheets used, and for a second case note to be entered when more than one exercise from different modules is used in an interview.

Table 5. Cross Validation of Case Note Content Identified by the C-QST against OIMS Module Information

OIMS Modules	Total No. Case Notes	Mod1 (%)	Mod2 (%)	Mod3 (%)	Mod4 (%)	Mod5 (%)	Mod6 (%)	Mod7 (%)	Mod8 (%)	Mod9 (%)	Mod10 (%)	Mod11 (%)	Mod12 (%)	Mod13 (%)	Unknown (%)	Mismatch (%)
PGI1	44193	84.1	10.7	0.1	0.2	0.2	0.2	0.4	0.1	0.1	0.1	0.1	0.1	0.5	15.0	0.9
PGI2	9766	3.8	71.1	3.2	0.3	0.3	0.2	0.5	0.2	0.1	0	0.4	0.2	0.7	24.9	4.0
PGI3	1252	1.8	1.8	64.1	0.7	1.0	0.3	0.9	0.1	0.3	0.1	0.3	0.2	1.3	32.2	3.7
PGI4	4577	15.5	9.7	1.6	74.2	1.1	0.2	0.7	0.3	0.1	0.3	0.2	0.2	2.0	21.4	4.4
PGI5	2548	2.5	1.5	0.4	2.6	79.2	1.1	0.8	0.2	0.1	0.2	0.4	0.2	2.1	15.3	5.5
PGI6	3019	10.1	0.6	0.1	0.4	3.3	65.7	0.9	0.1	0.1	0.1	0.3	0.2	1.5	28.6	5.7
PGI7	5086	1.9	12.3	0.1	0.2	0.4	1.5	74.7	0.1	0	0	0.1	0.2	1.3	22.6	2.7
PGI8	2325	13.4	0.9	1.7	0.3	0.4	1.6	1.1	74.3	0.2	1.4	0.1	1.0	1.2	20.1	5.6
PGI9	1189	1.8	1.8	0.2	0.3	1.4	0.3	0.3	4.9	70.5	1.9	0.2	0.1	2.0	21.3	8.2
PGI10	1046	2.3	0. 7	0	0.3	1.0	0.1	0.5	0.8	1.1	82.6	0.5	0.2	2.2	15.2	2.2
PGI11	2035	3.1	1.0	0.1	3.6	1.0	0.4	0.4	0.3	0.2	1.3	77.0	1.1	2.5	18.7	4.3
PGI12	1665	1.5	1.2	0.1	1.4	0.4	0.1	0.2	0.7	0.1	0.3	7.7	50.7	1.8	38.9	10.4
PGI13	3799	3.8	1.1	0.1	0.3	0.8	0.2	0.8	0.1	0.1	0	1.3	1.6	64.6	31.2	4.2

Table 6. The Proportion of Unknown Case Notes over Two Time Periods

OIMS Modules	June–Dec 2016	Jan–Aug 2017
PGI1	16.3	14.6
PGI2	31.1	24.2
PGI3	43.4	30.9
PGI4	28.3	20.4
PGI5	21.0	14.4
PGI6	40.0	27.1
PGI7	23.7	22.4
PGI8	31.3	18.8
PGI9	30.0	20.9
PGI10	29.7	13.9
PGI11	31.7	18.1
PGI12	54.1	37.5
PGI13	33.8	30.6

Exercise Level Information

Figure 2 shows that the most frequently used exercises were the mandatory ones that addressed initial assessment and case formulation (Exercises 1.1 and 1.2). A clearer view of the variation in exercise use can be seen in Figure 3, which replots Figure 2 without exercises from Modules 1 and 2. Among these non-mandatory exercises, it is clear that there is a large degree of variation. In particular, Exercise 9.4 ‘Communication Consequences’ was the least frequently used while Exercise 13.2 ‘Progress Review’ was the most frequently used. There also appears to be a distinctive pattern in which the first exercise within a module is the most frequently used exercise over the initial implementation period of the PGI.

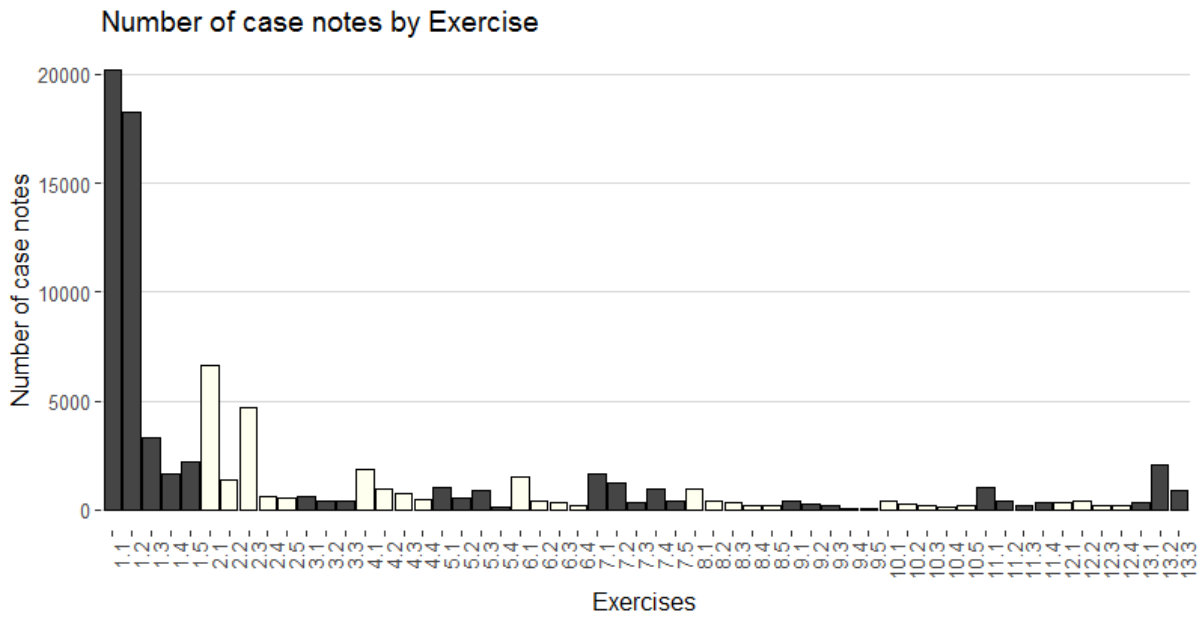

Figure 2. Distribution of Exercise Use across All PGI Modules (1–13)

Note: The alternating colour scheme was applied to aid visual discrimination of module membership.

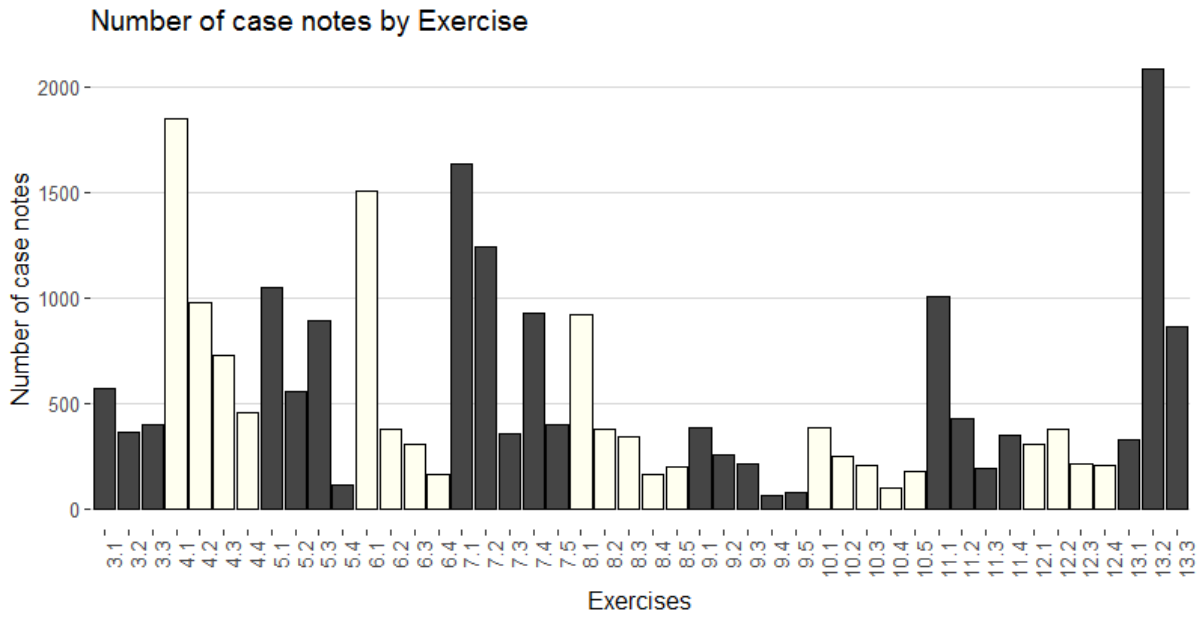

Figure 3. Distribution of Exercise Use across PGI Modules 3–13

Note: The alternating colour scheme was applied to aid visual discrimination of module membership.

Discussion

This paper presents the development of the C-QST and its utility in bridging the gap between qualitative and quantitative methods of data collection and interpretation. The C-QST is a tool that can automatically and reliably convert the content of large-scale unstructured case note text data into a structured and quantifiable form, providing access to valuable information latent in the text.

The tool adopts a step by step logic that utilises an a priori set of search terms (i.e., the PGI worksheets, exercise numbers and names, and module names) and applies a set of rules (i.e., flanking words preceding and following those search terms) to identify specific combinations of words following a pre-determined sequence. These search terms and rules are then refined over multiple iterations until an acceptable ratio of false positives to false negatives is achieved.

This approach, as demonstrated here, can be described as 'semi-conservative' in the language of signal detection theory (Harvey, 1992). This means that the C-QST could potentially overlook case notes with very low PGI specificity (e.g., case notes missing identified PGI keywords) despite being able to detect PGI related case notes with a high degree of accuracy. Therefore, the C-QST may be best conceptualised as a research tool that enables a researcher to sample and generate ad hoc datasets with cases in which there is clear evidence of PGI use.

Moreover, the ability of the C-QST to recognise exercise level information has clear benefits for process and outcome evaluation activities that require systematic data on relevant factors such as exercise type and dosage. The C-QST provides the researcher with the ability to define PGI use in a way that is both consistent and complies with their research needs across an entire stream of research studies.

Finally, it should be noted that there are significant time and cost savings from using the C-QST. A manual review of 50 case notes by one rater required approximately 20 minutes; therefore, a manual review of 82,500 case notes would require around 550 hours or three and a half months of full-time work. In contrast, once a set of rules have been defined, the C-QST requires only about three hours to batch process a set of 82,500 case notes.

As it is an automated tool, further savings are gained when the review process is repeated for future datasets. (PGI case notes are created on a daily basis and will continue to be created

over the operational lifespan of the intervention.) While the C-QST in its current form was designed to evaluate PGI related material, its logic can be extended to other contexts and it can be revised based on ad hoc research questions. This means that these time and cost benefits can be extended to other workstreams.

Limitations

The following limitations need to be considered when evaluating the present work. First, a caveat of the C-QST is that, for a PGI to be counted as a 'true delivery' by the script, case notes must contain explicit reference to the specific PGI module, exercise or worksheet being tested for. Any subtle or implicit references will not be recognised. As such, a proportion of the 'Unknowns' may be expected to comprise case notes that contain implicit references to valid PGI use, but which are not being detected by the tool. While this is possible, it must be noted that conservative rules were applied specifically to reject such cases. Acceptance of these ambiguous cases (in which the link with PGI is tenuous) would introduce a source of noise into the data due to the ambiguity inherent in these case notes. Moreover, if researchers—in the absence of any gold standard rule—were to make subjective judgements on how these ambiguous cases should be classified, bias could be introduced into the analysis. In this contest, it is important to note that the C-QST can be defined as liberally or conservatively as desired; as such, the rules can be changed to accommodate a more flexible and inclusive logic should this be required after the initial quality assurance phase.

Second, the flanking keyword rules are heuristic; while they represent a set of simple and efficient rules, they are neither exhaustive nor optimal. This means that they cannot account for all possible variations in misspelling, acronyms or contexts in which a PGI reference can be embedded. Therefore, the user needs to balance the trade-off between time spent accounting for possible and acceptable variations and the gain in sensitivity obtained. It is essential to clearly define a priori the parameters of the search, keeping in mind all possible limitations. It is expected that the accuracy of the tool will improve with closer alignment between C-QST search rules and criteria for recording case notes.

Third, the tool is highly sensitive to variations; a simple change in a rule or search term can potentially lead to significant differences in the results. For example, the C-QST currently makes no allowances for variations in exercise and worksheet names. This means that any word omission or changes in word sequence in names will not be recognised. A failure to

account for these variations may have a significant effect on the search hits. For example, further examination of Module 12, 'Pro-Social Lifestyle' (i.e., the module with the poorest agreement between the tool and OIMS category records) needs to be conducted to verify the appropriateness of the search terms used. Aside from that, the removal of any single flanking keyword will lead to a significant increase in false negatives. Therefore, the C-QST relies heavily on a user's forethought in defining the appropriate set of terms or rules in all stages of the process.

Finally, it needs to be noted that the C-QST is unable to reject negative statements such as '*did not* use Exercise 1.1'. This can be remedied in a future revision of the logic; however, in this example, there were very few instances of negative statements.

Conclusion

The C-QST is the first automated search function tool of its kind developed and used by CSNSW. With this demonstration, it is apparent that there is utility in the introduction of data triangulation and data mining techniques in CSNSW. A user may be required to conduct an initial qualitative review to define a set of initial parameters that can then be further refined through repeated runs of the C-QST. Once a set of parameters have been defined, the C-QST can be a powerful tool that enables researchers to use large qualitative datasets that are beyond manual ad hoc qualitative review.

References

Corrective Services New South Wales. (2016). *Community corrections practice guide for intervention handbook*.

Corrective Services New South Wales. (2017). *Guide to good case notes*.

Harvey, L. (1992). The critical operating characteristic and the evaluation of expert judgment. Organizational Behavior and Human Decision *Processes*, 53(2).

Landis, J. & Koch, G. (1977). The measurement of observer agreement for categorical data. B*iometrics*, 33(1).

Raudino, A., Neto, A. & Van Doorn, G. (2017). *Increase in the community corrections population. Corrections research, evaluation & statistics* (Research Digest No. 6). Retrieved from Corrective Services New South Wales http://www.correctiveservices.justice.nsw.gov.au/Documents/research-and-statistics/006-increase-community-corrections-population.pdf

R Core Team (2016). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.

Silge, J. & Robinson, D. (2016). Tidytext: Text mining and analysis using tidy data principles in R. *Journal of Open Source Software*, 1(3).

Appendix

Table A1. List of Regex Search Terms Used

Module	Search	Regex
Module 1	Module 1	assessment ^and\\b planning
	Exercise 1.1	^what\\b ^is\\b ^sup
		^getting\\b ^out\\b ^on\\b ^parole\\b,
		^sup ^expectation [:alnum:]
	Exercise 1.2	^offen ^map [:alnum:]
		^intervention\\b ^plan\\b
	Exercise 1.3	^cycle\\b ^of\\b ^change\\b
		readiness ^to\\b ^change\\b
	Exercise 1.4	decision balance chart
	Exercise 1.5	^impact\\b ^of\\b ^offen
Module 2	Module 2	^achieving\\b ^goal
	Exercise 2.1	^target\\b va.*l.*ue
		^achieving\\b ^goal
	Exercise 2.2	^immediate\\b ^gratifi [:alnum:]
		^problem\\b ^of\\b ^immediate\\b ^gratifi
	Exercise 2.3	^short ^and\\b ^long ^term\\b
		^step toward ^my\\b ^goal
	Exercise 2.4	^previous\\b ^suc.*s\\b
	Exercise 2.5	^starting\\b ^change\\b
Module 3	Module 3	dealing ^with\\b setback
	Exercise 3.1	defining ^suc.*s\\b
	Exercise 3.2	redefining failure
		redefining mistake
	Exercise 3.3	mistake ^map
Module 4	Module 4	manag.*ng stress\\b ^and\\b anger
	Exercise 4.1	recognising stress\\b
		stress\\b factor
		i.*den.*f.*ng stress\\b ^factor
	Exercise 4.2	^anger\\b volcano [:alnum:]
		^anger\\b ^word
		stress\\b ^and\\b anger diary

Module	Search	Regex
		^different ^level ^of\\b ^anger\\b
	Exercise 4.3	reacting ^to\\b stress\\b
	Exercise 4.4	reducing stress\\b
Module 5	Module 5	manag.*ng impulsitivity
	Exercise 5.1	planning ^ahead\\b
	Exercise 5.2	food ^for\\b thought
		acting without thinking
	Exercise 5.3	immediate reaction scenario
		stop\\b ^and\\b think\\b
		stop\\b think\\b act\\b
	Exercise 5.4	assumption ^and\\b belief
		belief ^and\\b assumption
Module 6	Module 6	manag.*ng environment
	Exercise 6.1	i.*den.*f.*ng ^high\\b ^risk\\b situation
		managing ^high\\b ^risk\\b situation
		i.*den.*f.*ng ^hrs\\b [:alnum:]
		managing ^hrs\\b [:alnum:]
	Exercise 6.2	i.*den.*f.*ng ^high\\b ^risk\\b people
	Exercise 6.3	saying ^no\\b
		avoiding ^high\\b ^risk\\b people
	Exercise 6.4	finding alternative
		managing ^high\\b ^risk\\b people
Module 7	Module 7	manag.*ng craving
	Exercise 7.1	^early warning ^sign
	Exercise 7.2	recognising craving
		craving log [:alnum:]
		recognising trigger for craving
	Exercise 7.3	coping ^with\\b craving
	Exercise 7.4	relapse prevention
		reducing ^risk\\b relapse
	Exercise 7.5	^lapse plan\\b
Module 8	Module 8	interpersonal relationship
	Exercise 8.1	people ^in\\b ^my\\b ^life\\b

Module	Search	Regex
		mapping relationship
	Exercise 8.2	relationship evaluation
		relationship cost benefit
	Exercise 8.3	relationship health check\\b
	Exercise 8.4	relationship belief
		relationship belief system
	Exercise 8.5	building ^a\\b good relationship
Module 9	Module 9	communication
	Exercise 9.1	^how\\b ^do\\b ^you\\b ^communicate\\b
		communication skill
	Exercise 9.2	communication barrier
		barrier ^to\\b effective communication
	Exercise 9.3	communication style
		assertive communication
	Exercise 9.4	effect ^of\\b communication style
		communication consequence
	Exercise 9.5	practicing ^asserti
		practicing ^asserti ^behavi
Module 10	Module 10	conflict resolution
	Exercise 10.1	^why\\b ^does conflict exist\\b
		purpose ^of\\b conflict
	Exercise 10.2	other point ^of\\b ^view
	Exercise 10.3	resolving conflict ^in\\b relationship
	Exercise 10.4	conflict resolution ^plan
		avoiding escalation ^of\\b conflict
	Exercise 10.5	^fair fighting [:alnum:]
		^rule ^for\\b ^fair fighting
Module 11	Module 11	^self\\b awareness [:alnum:]
	Exercise 11.1	my\\b strength ^and\\b skill ^my\\b strength
	Exercise 11.2	thought stopping [:alnum:]
		controlling thought [:alnum:]
	Exercise 11.3	thinking about thinking
	Exercise 11.4	^day\\b ^in\\b ^the\\b life

Module	Search	Regex
		weekly diary record
		awareness ^of\\b daily ^activit
Module 12	Module 12	prosocial lifestyle
		^pro\\b ^social\\b lifestyle
	Exercise 12.1	^past\\b prosocial relationship
		^past\\b ^pro\\b ^social\\b relationship
	Exercise 12.2	new\\b prosocial relationship
		^new\\b ^pro\\b ^social\\b relationship
		meeting ^new\\b people
	Exercise 12.3	belonging ^to\\b ^a\\b ^community
		va.*l.*ue ^in\\b ^common\\b
	Exercise 12.4	achievement ^plan\\b
		plan prosocial ^activit
		plan ^pro\\b ^social\\b ^activit
Module 13	Module 13	general skill
	Exercise 13.1	problem solving ^plan\\b
	Exercise 13.2	^progress\\b ^review\\b
	Exercise 13.3	^mindful [:alnum:] [:alnum:]
		practicing ^mindful [:alnum:]

Corrections Research, Evaluation & Statistics
Governance and Continuous Improvement
Corrective Services NSW
GPO Box 31
Sydney NSW Australia

Telephone (02) 8346 1556
Facsimile (02) 8346 1590
Email: research.enquiries@justice.nsw.gov.au

Material published by
Corrections Research, Evaluation & Statistics includes:

Research Publications
Research Bulletins
Research Digests
Statistical Publications