User rating activity within KIWI: A technology for public health event monitoring and early warning signal detection

Ellie Andres[1], **Shamir Mukhi**[1*]

1. Canadian Network for Public Health Intelligence, National Microbiology Laboratory, Winnipeg, MB

Abstract

Objectives: To review user signal rating activity within the Canadian Network for Public Health Intelligence's (CNPHI's) Knowledge Integration using Web-based Intelligence (KIWI) technology by answering the following questions: (1) who is rating, (2) how are users rating, and (3) how well are users rating?

Methods: KIWI rating data was extracted from the CNPHI platform. Zoonotic & Emerging program signals with first rating occurring between January 1, 2016 and December 31, 2017 were included. Krippendorff's alpha was used to estimate inter-rater reliability between users. A z-test was used to identify whether users tended to rate within 95% confidence interval (versus outside) the average community rating.

Results: The 37 users who rated signals represented 20 organizations. 27.0% (n = 10) of users rated ≥10% of all rated signals, and their inter-rater reliability estimate was 72.4% (95% CI: 66.5-77.9%). Five users tended to rate significantly outside of the average community rating. An average user rated 58.4% of the time within the signal's 95% CI. All users who significantly rated within the average community rating rated outside the 95% CI at least once.

Discussion: A diverse community of raters participated in rating the signals. Krippendorff's Alpha estimate revealed moderate reliability for users who rated ≥10% of signals. It was observed that inter-rater reliability increased for users with more experience rating signals.

Conclusions: Diversity was observed between user ratings. It is hypothesized that rating diversity is influenced by differences in user expertise and experience, and that the number of times a user rates within and outside of a signal's 95% CI can be used as a proxy for user expertise. The introduction of a weighted rating algorithm within KIWI that takes this into consideration could be beneficial.

Keywords: Public health intelligence, digital disease detection, event monitoring, early warning, data mining, event-based surveillance.

*Correspondence: shamir.mukhi@canada.ca

DOI: 10.5210/ojphi.v10i2.8547

Background

The use of novel internet-based public health intelligence monitoring techniques for the purpose of providing early warning and situation awareness of potential health threats to support surveillance activities has grown tremendously over the last few decades [1-9]. Technologies have been reviewed and evaluated in the literature and authors have suggested a variety of directions for enhancements and growth opportunities. One of these directions is the need for establishing collaborative networks of public health professionals for the verification and dissemination of early warning signals [3]. To meet this need, the National Microbiology Laboratory's Canadian Network for Public Health Intelligence (CNPHI) developed an innovative technology for public health event monitoring and early warning signal detection called Knowledge Integration using Web-based Intelligence (KIWI) within the context of its existing platform hosting thousands of public health professionals and health-related communities [10]. The CNPHI platform was established in 2003 as an initiative of the Public Health Agency of Canada, and it provides a variety of scientific informatics tools within the framework of six focus areas: Knowledge Management, Collaboration, Surveillance, Alerting, Event Management, and Laboratory Systems. The KIWI technology aims to complement and support surveillance activities being performed on the CNPHI platform and is made available under Knowledge Management.

KIWI was designed to facilitate event monitoring and early warning signal detection for multiple types of public health related events. The Zoonotic & Emerging (ZE) program, for example, focuses on events related to zoonotic, emerging and re-emerging disease. The ZE program was customized for – and in collaboration with – the Centre for Emerging and Zoonotic Disease (CEZD) community, piloted in 2015 and initiated in 2016. CEZD has created an active community of professionals from various authorities using CNPHI's Collaboration Centre and KIWI technologies.

As a quick overview of KIWI, the technology collects individual information pieces (IIPs) from a selection of program-specific online sources, collated, and then analyzed using a program-specific sense making algorithm (SMA), which utilizes three dictionaries (hazards, relevant terms, and significant terms) to identify signals of interest. These three steps are performed through an automated process. IIPs identified with early warning potential are classified as Anticipatory Intelligence Signals (AISs) and are presented to all users for community rating. Users can rate each AIS on a Likert-type scale from 1 (Not Relevant) to 5 (Extremely Relevant). The purpose of user rating is to verify which AISs should become Early Warning Signals (EWSs) based on a pre-defined rating value threshold. The resulting EWSs can be summarized as reports and used to support public health event monitoring and surveillance activities by providing synthesized intelligence and situational awareness.

A key feature of the KIWI technology is that there is a community of program-specific experts involved in the rating of AISs. The CNPHI platform is used by a variety of public health professionals from multiple jurisdictions, disciplines, and areas of expertise. The purpose of this analysis is to answer the following questions: (1) who is rating, (2) how are users rating and (3) how well are users rating? These questions will be answered within the context of KIWI's ZE program.

Methods

KIWI rating data was extracted from the CNPHI platform on January 13, 2018. The dataset contained observations from June 1, 2015 to January 12, 2018 and included the following variables: program ID, signal ID, user ID, rating, and rating date/time. The dataset was refined in SPSS using two primary exclusion criteria: (a) data linked to signals with first user rating occurring before 2016 or after 2017, and (b) data linked to programs other than the ZE program.

Who is rating?

User organizations for those participating in KIWI signal rating were extracted using KIWI's analytics feature for the time period of January 1, 2016 to December 31, 2017. Two variables "jurisdiction" and "type of organization" were created to broadly describe the organizations represented by these raters.

How Are Users Rating?

The following variables were derived from the original data (see Table 1):

- *Year of first rating per signal:* the year in which the first rating occurred for a given signal;
- *Number of signals rated per user:* total count of signals rated by a specific user;
- *Grouped proportion of signals rated per user:* Users were grouped based on their proportion of signals rated: (1) Less than 10%, (2) 10% to 90%, and (3) More than 90%;
- *Number of days rated per user:* total number of days a specific user rated signals;
- *Rating duration per signal:* total duration in days for all ratings for a given signal;
- *Number of signals rated per day per user:* count of signals rated by a specific user per day;
- *Day of the week:* specific day of the week the rating for a specific signal took place;
- *Number of users rating per signal:* total count of users rating specific signals;
- *Average and median rating per signal:* mean and median values for all ratings per signal.

Table 1: Derived variables for analysis.

Derived Variable	Base Variable(s)	Formula/Logic	SPSS
Year of first rating per signal	o signal ID o date/time	Sorted data by signal ID and date/time. Identified first rating per signal and extracted year from date/time.	o IDENTIFY DUPLICATES o XDATE.YEAR
Number of signals rated per user	o signal ID o user ID	Sorted data by user ID and signal ID. Count of signal ID per user.* *Note: signals can only be rated once per user.	o AGGREGATE

Grouped proportion of signals rated per user	o signal ID o user ID	Formula = a/b. (a) Numerator: number of signals rated per user.* (b) Denominator: total number of signals. *Note: signals can only be rated once per user.	o AGGREGATE o IDENTIFY DUPLICATES o COMPUTE VARIABLE
Number of days rated per user	o user ID o date/time	Extracted date from date/time. Sorted data by user ID and date. Count of rating date/time per user ID.	o XDATE.DATE o IDENTIFY DUPLICATES o AGGREGATE
Rating duration per signal	o signal ID o date/time	Extracted date from date/time. Sorted data by signal ID and date. Identified both first and last ratings by date. Calculated the difference in days plus one day. Formula = (date of last rating – date of first rating) + 1	o XDATE.DATE o IDENTIFY DUPLICATES o COMPUTE VARIABLE
Number of signals rated per day per user	o signal ID o user ID o date/time	Extracted date from date/time. Sorted data by user ID, date, and signal ID. Count of signals per day per user.	o XDATE.DATE o AGGREGATE
Day of the week	o date/time	Extracted day of the week using XDATE.WKDAY command.	o XDATE.WKDAY
Number of users rating per signal	o signal ID o user ID	Sorted data by signal ID and user ID. Count of user ID per signal.	o AGGREGATE
Average and median rating per signal	o signal ID o rating	Sorted data by signal ID. Mean and median of rating per signal.	o AGGREGATE

How Well Are Users Rating?

Inter-rater reliability

Inter-rater reliability is used to reflect the variation between two or more raters who measure the same group of signals [11]. The intra-class correlation coefficient (ICC) is commonly used to assess inter-rater reliability; however, it cannot accommodate datasets with a large number of missing data due to its use of list-wise deletion [12]. For example, our dataset has 37 users, but not all users have rated each signal to produce a fully-crossed design. Therefore, a huge proportion of our data would be dropped during the calculation of ICC. A more appropriate method, in our case, is the use of Krippendorff's alpha, which is capable of handling missing data [12,13]. In response to the call for a standard reliability measure, Andrew F. Hayes and Klaus Krippendorff proposed in 2007 that Krippendorff's alpha meets the criteria for a good index of reliability, can be used for any number of raters, levels of measurement, and sample sizes, and can be used in the presence or absence of missing data [13]. Hayes developed a SPSS macro/code for calculating Krippendorff's alpha called "KALPHA" [14]. Krippendorff's alpha was calculated using an ordinal measurement level with all users, and again with users grouped by the proportion of signals rated per user.

User rating and the community norm

Average ratings and 95% confidence intervals (CI) were calculated for each signal with more than one rating. Individual user ratings were then compared to the average rating per signal. If the user's rating was within the 95% CI of the signal's average rating, then it was grouped as "within" and if it was outside of the 95% CI of the signal's average rating, then it was grouped as "outside"; thus, creating a dichotomous variable. Those "outside" were further identified as either "above" or "below" depending on whether the individual rating was greater than the signal's average upper limit or less than signal's average lower limit, respectively (Figure 1).

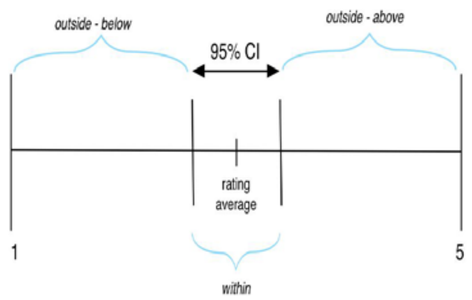

Figure 1: Signal ratings within and outside of 95% CI.

A one-sample z-test was used to identify whether the proportion of user ratings classified as "within" (rather than "outside") was significantly different than 50% (i.e. no difference). The z-statistic was calculated using the formula:

$$Z= \frac{(\hat{p} - p_o)}{\sqrt{\frac{p_o*(1-p_o)}{n}}}$$

$\hat{p}$ is the observed proportion of "within" ratings, p_o is the expected proportion of "within" ratings, and n is the total number of ratings. The null hypothesis was rejected if $1.645 < Z < -1.645$, which is to say if the p-value was greater than 0.05.

Results

Who is rating?

During 2016 and 2017, there were 37 CNPHI users participating in KIWI signal rating within the ZE program. These 37 users represented 20 unique organizations. User organizations stemmed from government (federal and provincial), industry, academia, and professional groups (Figure 2). Jurisdictionally, users represented international, national (Canada), and provincial authorities. The most highly represented jurisdiction was national with 21 users. Of these 21 users, 81.0% (n = 17) represented a government organization and 19.0% (n = 4) represented a professional group. The second leading jurisdiction included the provinces with 15 users. Of these 15 users, 60.0% (n = 9) represented a government organization, 13.3% (n = 2) represented industry, and 26.7% (n = 4) represented an academic organization.

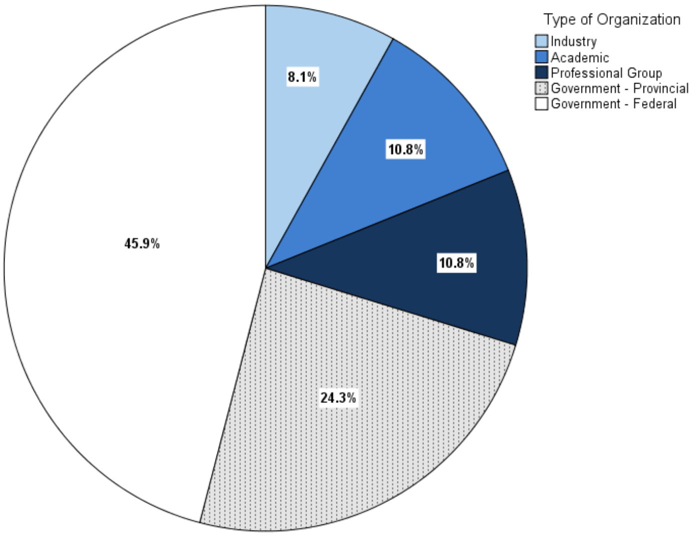

Figure 2: Types of organizations represented by users participating in KIWI signal rating within the ZE program during 2016-2017.

How Are Users Rating?

A total of 6988 signals were rated within the ZE program during 2016 and 2017. Of these signals, 48.7% (n = 3400) were first rated in 2016, and 51.3% (n = 3588) were first rated in 2017.

Based on grouped proportion of signals rated per user:

- *Less than 10%:* 73.0% (n = 27) of users rated less than 10% of all rated signals. The number of signals rated by these 27 users ranged from 1 to 609 signals with an average of 88 signals, median of 12 signals, and mode of 1 signal. The number of days in which rating took place ranged from 1 to 51 days with an average of 7 days, median of 2 days, and mode of 1 day. Of the 27 users, 22.2% (n = 6) rated in 2016 only, 29.6% (n = 8) rated in both 2016 and 2017, and 48.1% (n = 13) rated in 2017 only.

- *10% to 90%:* 21.6% (n = 8) of users rated 10% to 90% of all rated signals. The number of signals rated by these 8 users ranged from 978 to 5698 signals with an average of 3122 signals and median of 2896 signals. The number of days in which rating took place ranged from 33 to 345 days with an average of 149 days and median of 129 days. Of the 8 users, 12.5% (n = 1) rated in 2016 only and 87.5% (n = 7) rated in both 2016 and 2017.
- *More than 90%:* 5.4% (n = 2) of users rated more than 90% of all rated signals. The number of signals rated by these 2 users ranged from 6323 to 6325 signals with an average and median of 6324 signals. The number of days in which rating took place ranged from 529 to 587 days with an average and a median of 558 days. Both users rated in both 2016 and 2017.

In total, the number of signals rated per user ranged from 1 to 6325 signals with an average of 1081 signals, median of 57 signals, and mode of 1 signal. The number of days in which rating took place ranged from 1 to 587 days with an average of 68 days, median of 5 days, and mode of 1 day. Of the 37 users, 18.9% (n = 7) rated in 2016 only, 46.0% (n = 17) rated in both 2016 and 2017, and 35.1% (n = 13) rated in 2017 only. A summary of user activity is displayed in Figure 3.

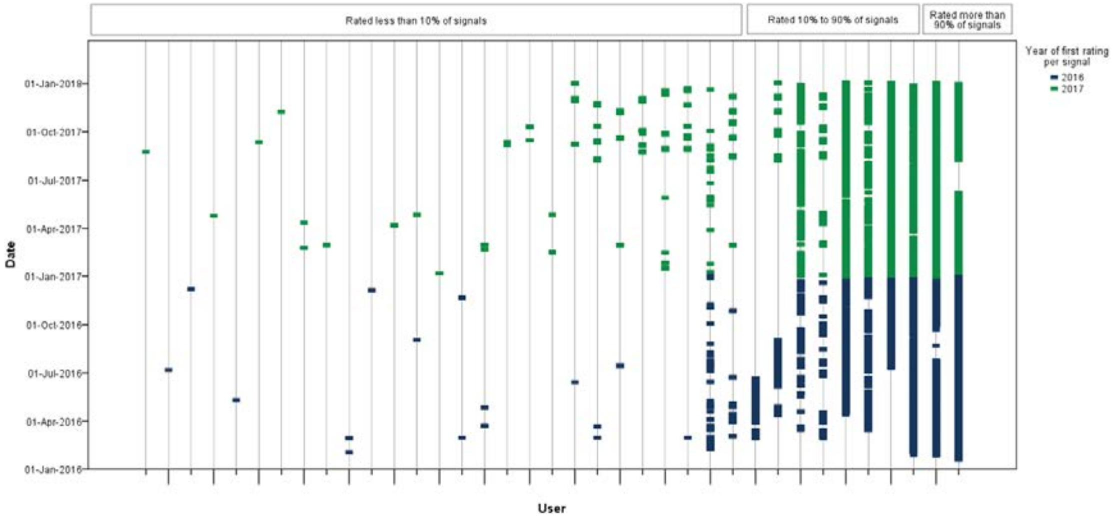

Figure 3: User rating activity during 2016 and 2017 within KIWI's ZE program.

As a community, 50% of signals were rated within 1-5 days of being identified by the technology, 95% were rated within 1-8 days, and 99% were rated within 1-19 days. Overall, the community rated within 1-180 days with an average of 6 days, median of 5 days, and mode of 7 days. The number of signals rated per day ranged from 1 to 102 with an average of 31, median of 23, and mode of 12. When results were separated by day of the week, the average number of signals was 19 with a median of 17 for Saturday and Sunday, 64 with a median of 67 on Monday, and 28 with a median of 22 for the remaining weekdays (Figure 4). The number of users rating per signal was normally distributed (Figure 5) with an average, median, and mode of 6 users and range of 1-11 users rating per signal. An average signal was rated as 1.7 with a median rating of 2 (Some Relevance). An overview of community signal rating is provided in Table 2.

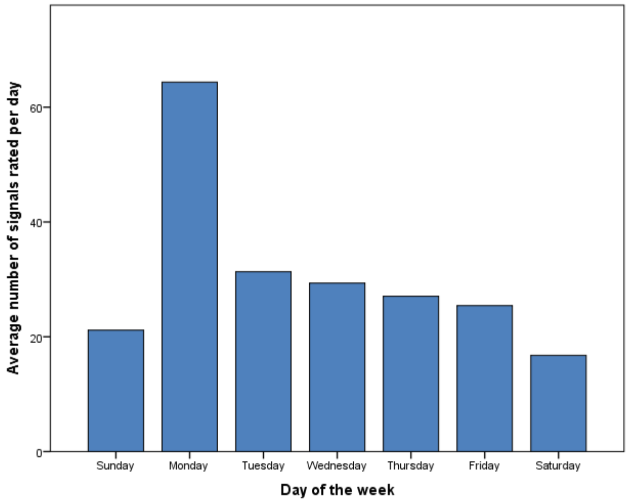

Figure 4: Average number of signals rated per day by day of the week.

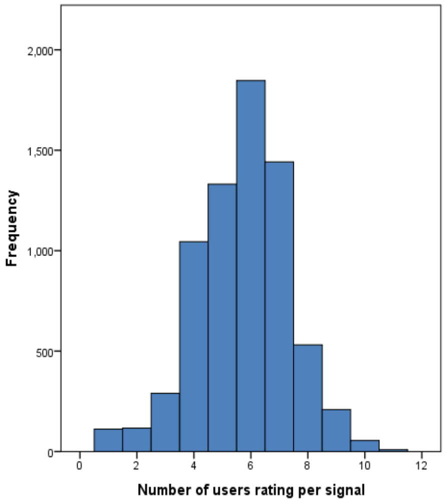

Figure 5: Distribution of the number of user ratings per signal.

Table 2: An overview of community signal rating within KIWI's ZE program during 2016-2017.

	Average Rating	# of Signals	% of Signals
1 – Not Relevant	$\bar{x} < 1.5$	2897	41.5
2 – Some Relevance	$1.5 \leq \bar{x} < 2.5$	3151	45.1
3 – Relevant	$2.5 \leq \bar{x} < 3.5$	923	13.2
4 – Very Relevant	$3.5 \leq \bar{x} < 4.5$	17	0.2
5 – Extremely Relevant	$\bar{x} \geq 4.5$	0	0.0
Total	*1-5*	*6988*	*100.0*

How Well Are Users Rating?

Inter-rater reliability

Overall, Krippendorff's Alpha reliability estimate was 69.6% (95% CI: 64.5-74.4%). When users were grouped by their proportion of signals rated, the reliability estimate increased with an increase in the proportion of signals rated. Those who rated less than 10% of signals had an estimate of 42.1% (95% CI: 33.8-49.8%), those who rated 10% to 90% had an estimate of 70.1% (95% CI: 63.9-75.9%), and those who rated more than 90% had an estimate of 77.1% (95% CI: 71.7-82.0%). Combining the last two groups, the reliability estimate for users who rated greater than or equal to 10% of signals was 72.4% (95% CI: 66.5-77.9%).

User rating and the community norm

Of the 6988 rated signals, 98.4% (n = 6876) were rated by more than one user. For these 6876 signals, an average user rated 58.5% of the time within the signal's 95% CI (95% CI: 49.5-67.2%). Comparing the proportion of signals rated within the signal's 95% CI per user to 50% (i.e. no difference), it was found that the raters can be categorized into three groups as follows (see Figure 6):

Category I: No significant result: 29.7% (11/37) of users did not produce a significant result.

Category II: Significant result and significantly lower than 50% of signals within 95% CI: 13.5% (5/37) of users produced a significant result and had a negative z-score. These users rated significantly less than 50% of signals within the signal's 95% CI. For these users, the proportion of signals rated outside and above the 95% CI was compared to 50% (i.e. no difference), and it was found that one user did not produce a significant result.

Category III: Significant result and significantly higher than 50% of signals within 95% CI: 56.8% (21/37) of users produced a significant result and had a positive z-score. These users rated significantly greater than 50% of signals within the signal's 95% CI.

In summary, there were 5 users who rated outside of the average community rating. Of these 5 users, three tend to rate high, one tends to rate low, and one has no significant tendency. Each of these 5 users rated less than 10% of signals. Important to note that all users who rate significantly greater than 50% of signals within their 95% CIs have rated outside of the 95% CI at least once.

Discussion

Users participating in rating KIWI signals within the ZE program are multi-jurisdictional and represent multiple types of organizations. The community is well represented; however, the majority of users were not regularly rating signals. It appears that approximately one in four users rated greater than or equal to 10% of signals over the past 2 years. Another distinction between users is whether the user was newly active or had stopped being active. For instance, 35.1% of users were active in both 2016 and 2017, but 18.9% of users stopped rating in 2016 and 46.0% of users only began rating in 2017.

Krippendorff's Alpha estimate revealed moderate reliability for users who rated greater than or equal to 10% of signals. A Krippendorff's Alpha estimate of ≥80% is considered good reliability, and an estimate of ≤67% is considered low reliability [15]. Overall, 5 users were found to rate significantly below average; however, these users rated less than 10% of signals and therefore did not contribute to this particular estimate.

Since the overall reliability estimate was poor for all users, it could be insightful to review the use of rating weights per user based on the user's level of expertise in the field relevant to the program. Evidence suggests that there is an increase in reliability estimate for users who rated an increased proportion of signals. An additional inquiry is whether the frequency of signal rating per user could be used as a proxy approach for determining user rating weights. A limitation of this proxy method is that the reliability estimate indicates the degree of user agreement or similarity but is not necessarily indicative of ‘truth’. Development of user profiles to obtain measures on potential parameters that could be incorporated in a new weighted rating algorithm could be helpful. Profile parameters could include: *Expertise*: hazards the user has an expert level of knowledge on, and *Experience*: rating frequency.

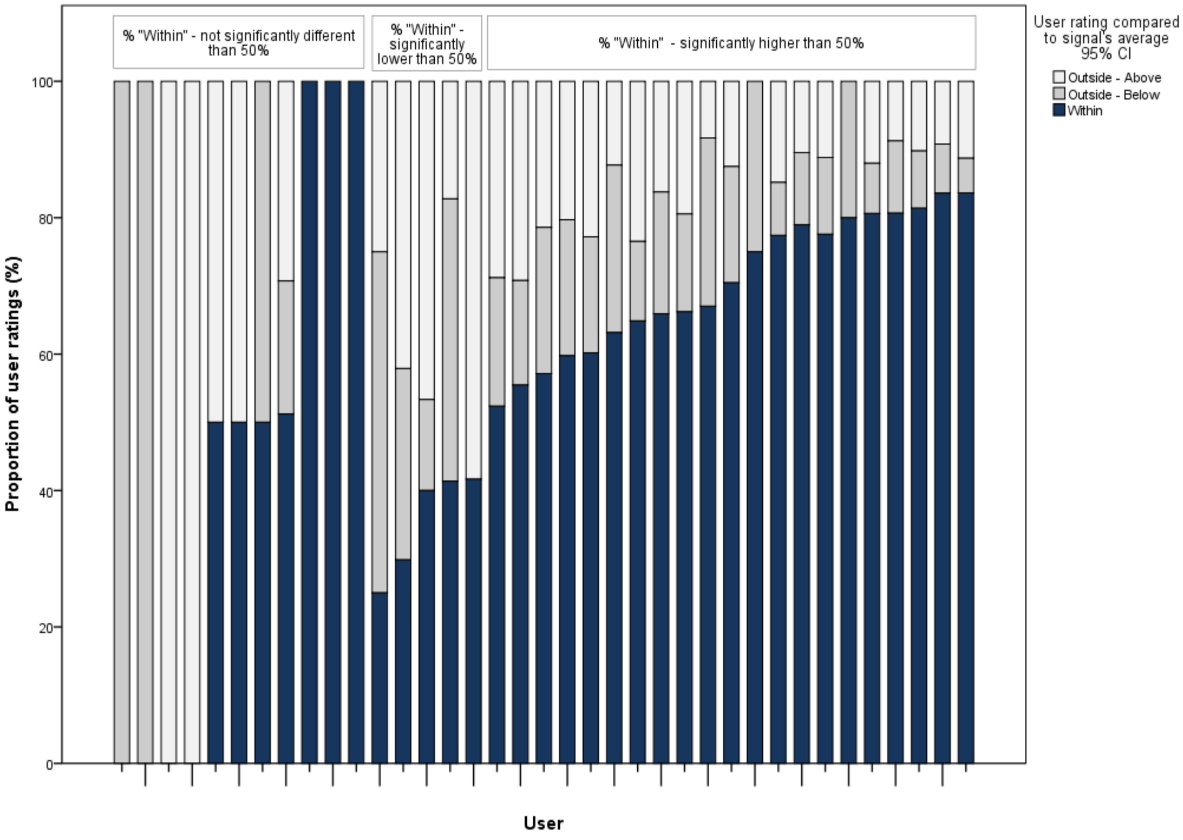

Figure 6: The proportion of signals rated within, outside and above, and outside and below the average 95% CI per signal by user.

Conclusions

A diverse community of users (n = 37) participated in rating KIWI signals within the Zoonotic & Emerging program over a two-year period (2016-2017). Users represented multiple jurisdictions (i.e. provincial, national, and international) and types of organizations (i.e. industry, academia, government, and professional groups). Users who rated at least 10% of all rated signals produced a moderate inter-rater reliability estimate; however, poor reliability was found among users when all raters were included in the calculation. Results indicate that diversity remains between user ratings. It is hypothesized that differences in user rating may be due to varying levels of interest or expertise among raters regarding the hazard represented within each signal. Inter-rater reliability estimates were compared by the proportion of signals rated, and it was observed that estimates improved with an increase in user’s experience in rating signals. An additional hypothesis is that number of times the user rates within and outside (above or below) a signal’s 95% CI can be used as a proxy for user expertise. The introduction and further analysis of a weighted rating algorithm within the KIWI technology that takes this into consideration could be beneficial.

Limitations

Inter-rater reliability is an estimate of rating similarity rather than an estimate of truth. This must be taken into account during result interpretation. Sample size related to the number of users participating in rating activities was good overall (n = 37). However, only 27.0% of users (n = 10) rated greater than or equal to 10% of all rated signals, and as a subset of this group, only 2 users rated more than 90% of all rated signals. Average and median results for this particular group should be interpreted with caution. Inter-rater reliability estimates were not dependent on rater sample size, rather signal sample size – which was very large. With this in mind, estimates became more precise moving from users who rated less than 10% of signals to those who rated more than 90% of signals.

Acknowledgements

We would like to acknowledge that the Zoonotic & Emerging program was developed in collaboration with the CEZD community and enabled through the CEZD-IIR project, which was funded by the Canadian Safety and Security Program, managed through Defence Research and Development Canada's Centre for Security Science, and hosted by the Canadian Food Inspection Agency. Authors would like to acknowledge the CNPHI team, the CEZD project team, and the CEZD rating community.

References

1. Hartley DM, Nelson NP, Arthur R, Barboza P, Collier N, et al. 2013. An overview of internet biosurveillance. *Clin Microbiol Infect*. 19(11), 1006-13. PubMed https://doi.org/10.1111/1469-0691.12273

2. Dion M, AbdelMalik P, Mawudeku A. 2015. Big data and the global public health intelligence network (GPHIN). *Can Commun Dis Rep*. 41(9), 209. PubMed https://doi.org/10.14745/ccdr.v41i09a02

3. Keller M, Blench M, Tolentino H, Freifeld CC, Mandl KD, et al. 2009. Use of unstructured event-based reports for global infectious disease surveillance. *Emerg Infect Dis*. 15(5), 689-95. PubMed https://doi.org/10.3201/eid1505.081114

4. Hartley DM, Nelson NP, Walters R, Arthur R, Yangarber R, et al. 2010. The landscape of international event-based biosurveillance. *Emerg Health Threats J*. 3(e3). PubMed https://doi.org/10.3402/ehtj.v3i0.7096

5. Lyon A, Nunn M, Grossel G, Burgman M. 2012. Comparison of Web-Based biosecurity intelligence systems: BioCaster, EpiSPIDER and HealthMap. *Transbound Emerg Dis*. 59(3), 223-32. PubMed https://doi.org/10.1111/j.1865-1682.2011.01258.x

6. Barboza P, Vaillant L, Mawudeku A, Nelson NP, Hartley DM, et al. 2013. Evaluation of epidemic intelligence systems integrated in the early alerting and reporting project for the detection of A/H5N1 influenza events. *PLoS One*. 8(3), e57252. PubMed https://doi.org/10.1371/journal.pone.0057252

7. Barboza P, Vaillant L, Le Strat Y, Hartley DM, Nelson NP, et al. 2014. Factors influencing performance of internet-based biosurveillance systems used in epidemic intelligence for early detection of infectious diseases outbreaks. *PLoS One*. 9(3), e90536. PubMed https://doi.org/10.1371/journal.pone.0090536

8. Velasco E, Agheneza T, Denecke K, Kirchner G, Eckmanns T. 2014. Social media and Internet-Based data in global systems for public health surveillance: A systematic review. *Milbank Q*. 92(1), 7-33. PubMed https://doi.org/10.1111/1468-0009.12038

9. O'Shea J. 2017. Digital disease detection: A systematic review of event-based internet biosurveillance systems. *Int J Med Inform*. 101, 15-22. PubMed https://doi.org/10.1016/j.ijmedinf.2017.01.019

10. Mukhi SN, Andres E, Demianyk B, Gammon B, Kloeze H. 2016. KIWI: A technology for public health event monitoring and early warning signal detection. *Online J Public Health Inform*. 8(3), e208. PubMed https://doi.org/10.5210/ojphi.v8i3.6937

11. Koo TK, Li MY. 2016. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med*. 15(2), 155-63. PubMed https://doi.org/10.1016/j.jcm.2016.02.012

12. Hallgren KA. 2012. Computing inter-rater reliability for observational data: an overview and tutorial. *Tutor Quant Methods Psychol*. 8(1), 23. PubMed https://doi.org/10.20982/tqmp.08.1.p023

13. Hayes AF, Krippendorff K. 2007. Answering the call for a standard reliability measure for coding data. *Commun Methods Meas*. 1(1), 77-89. https://doi.org/10.1080/19312450709336664

14. Hayes AF. 2012. My Macros and Code for SPSS and SAS - KALPHA. Available at: http://afhayes.com/spss-sas-and-mplus-macros-and-code.html

15. Krippendorff K. Content analysis: An introduction to its methodology. 2nd ed. USA: Sage; 2012.

16. Sullivan GM, Artino AR, Jr. 2013. Analyzing and interpreting data from Likert-type scales. *J Grad Med Educ*. 5(4), 541-42. PubMed https://doi.org/10.4300/JGME-5-4-18