JAMIA Open, 00(0), 2022, 1–2
doi: 10.1093/jamiaopen/ooab108
Corrigendum

Corrigendum

Corrigendum to: Enabling qualitative research data sharing using a natural language processing pipeline for deidentification: moving beyond HIPAA Safe Harbor identifiers

Aditi Gupta, Albert Lai, Jessica Mozersky, Xiaoteng Ma, Heidi Walsh, James M DuBois

JAMIA Open, Volume 4, Issue 3, July 2021, ooab069, https://doi.org/10.1093/jamiaopen/ooab069

In the originally published version of this manuscript, there were inaccuracies in the frequencies of identifiers. In fourth paragraph of **RESULTS,** the following should read: "The number of identifiers varied in our 2 datasets but ranged between 1-3% of the total word tokens (Table 3). These identifiers are located within large volumes of unstructured qualitative text, making manually locating them challenging." instead of: "The number of identifiers varied in our 2 datasets but ranged between 0.01% and 0.03% of the total word tokens (Table 3). The overall frequency of identifiers is relatively low due to the large volume of unstructured qualitative text within which identifiers are located." And in the second paragraph of **DISCUSSION**, the following should read: "We found very few HSH and non-HSH identifiers in qualitative text (1-3% of all words)." instead of "We found very few HSH and non-HSH identifiers in qualitative text (0.01-0.03% of all words)." Frequencies were affected in the Identifier Count column of Table 3 also. This should read:

Table 3. Descriptive statistics of the 2 datasets used in the study and the number (%) of identifiers (HSH and non-HSH) extracted using the NLP pipeline from each set; and gold-standard evaluation of the NLP system

	Dataset name (number of files)	Token count	Identifier count (%)	Precision	Recall	F1 score
Pilot files	NIB stories (6 files)	12 620	389 (3%)	0.93	0.92	0.93
	QDS interviews (9 files)—Iteration 1	85 590	650 (1%)	0.98	0.83	0.90
	QDS interviews (9 files)—Iteration 2[a]	85 590	650 (1%)	0.98	0.90	0.94
Additional files	NIB stories (25 files)	48 807	858 (2%)	0.93	0.98	0.95
	QDS interviews (30 files)—Iteration 1	139 323	998 (1%)	0.97	0.81	0.88
	QDS interviews (30 files)—Iteration 2[a]	139 323	998 (1%)	0.97	0.95	0.96
Total	70	286 340	2888 (1%)	0.95	0.88	0.91
Total—Iteration 2[a]	70	286 340	2888 (1%)	0.95	0.96	0.96

[a]We performed an error analysis after Iteration 1 and observed that a single name of the organization which repeated as part of an interview question in every transcript of dataset 2, was being missed by our pipeline and driving the low recall. Iteration 2 results show the performance of the pipeline after the removal of one problematic organization name that was not recognized.

HSH: HIPAA Safe Harbor; NIB: *Narrative Inquiry in Bioethics*; QDS: qualitative data sharing.

Instead of

Table 3. Descriptive statistics of the 2 datasets used in the study and the number (%) of identifiers (HSH and non-HSH) extracted using the NLP pipeline from each set; and gold-standard evaluation of the NLP system

	Dataset name (number of files)	Token count	Identifier count (%)	Precision	Recall	F1 score
Pilot files	NIB stories (6 files)	12 620	389 (0.03 %)	0.93	0.92	0.93
	QDS interviews (9 files)—Iteration 1	85 590	650 (0.01 %)	0.98	0.83	0.90
	QDS interviews (9 files)—Iteration 2[a]	85 590	650 (0.01 %)	0.98	0.90	0.94
Additional files	NIB stories (25 files)	48 807	858 (0.02%)	0.93	0.98	0.95
	QDS interviews (30 files)—Iteration 1	139 323	998 (0.01%)	0.97	0.81	0.88
	QDS interviews (30 files)—Iteration 2[a]	139 323	998 (0.01%)	0.97	0.95	0.96
Total	70	286 340	2888 (0.01%)	0.95	0.88	0.91
Total—Iteration 2[a]	70	286 340	2888 (0.01%)	0.95	0.96	0.96

[a]We performed an error analysis after Iteration 1 and observed that a single name of the organization which repeated as part of an interview question in every transcript of dataset 2, was being missed by our pipeline and driving the low recall. Iteration 2 results show the performance of the pipeline after the removal of one problematic organization name that was not recognized.

HSH: HIPAA Safe Harbor; NIB: *Narrative Inquiry in Bioethics*; QDS: qualitative data sharing.

These inaccuracies have now been corrected online.