**Appendix 1.**

We applied natural language processing (NLP) to extract different types of polyps from pathology reports. Our algorithm includes an iterative process of text preprocessing and application of regular expression. In our text preprocessing, we changed all words to uppercase and removed all stopwords and punctuations. In regular expression, we used a series of technique by extracting of keyword “polyp” and its location, search for polyp subtype (“tubular”, “tubulovillous”, “villous”, etc) adjacent to our keyword. Subsequently, we compared negation terms within range of our keywords to determine if polyp is negative. Schematic diaphragm below illustrates our process.

A diagram of a flowchart

Description automatically generated

We split our pathology reports into several training and validation datasets, iterating step 1-5 until we optimized our accuracy. We subsequently used a test dataset containing 1000 samples to determine our accuracy. Authors ESH and SYL clinically reviewed these reports to match to our NLP algorithm. Our overall adenoma agreement was 98.0% (ƙ=0.959).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | | **NLP** | | |
| 0 | 1 | Total |
| **Clinical Review** | 0 | 418 | 17 | 435 |
| 1 | 3 | 562 | 565 |
| Total | 421 | 579 | 1000 |
|  | | | ƙ | % |
| 0.959 | 98% |

The agreement did vary by adenoma subtypes (94.3%, ƙ=0.971 for tubular; 96%, ƙ=0.981 for tubulovillous; 44.2%, ƙ=0.665 for villous; 96%, ƙ=0.980 for sessile serrated)

**Supplementary Table 1. Pathology Reports**

|  |  |
| --- | --- |
| **Pathology1** | **N=88,091 (%)** |
| Adenoma | 55,524 (63.0) |
| Tubular adenoma | 54,818 (62.2) |
| Tubulovillous adenoma | 1,812 (2.1) |
| Villous adenoma | 350 (0.4) |
| Sessile serrated lesions | 12,177 (13.8) |

15478 pathology had both adenoma and sessile serrated lesions