Breaching the curation bottleneck with human-machine reading symbiosis

Cliff Wong¹, Rajesh Rao¹, Taofei Yin², Cara Statz², Susan Mockus³, Sara Patterson²,
Hoifung Poon¹

¹Microsoft Research, Redmond, WA, USA

²The Jackson Laboratory for Genomic Medicine, Farmington, CT, USA

³Precision Biomarker Laboratories, Cedars-Sinai, Los Angeles, CA, USA

Corresponding authors:

Sara Patterson

10 Discovery Drive, The Jackson Laboratory for Genomic Medicine, Farmington, CT,
USA 06032 (Phone) 860-837-2382 (E-mail) sara.patterson@jax.org

Hoifung Poon

One Microsoft Way, Microsoft Research, Redmond, WA 98052, USA (Phone) 425-421-
4791 (E-mail) hoifung@microsoft.com

Disclaimer:

None

Abstract

Purpose

The explosion of molecular biomarker and treatment information in the precision medicine era drastically exacerbated difficulty in identifying patient-relevant knowledge for clinical researchers and practitioners. Curated knowledgebases, such as the JAX Clinical Knowledgebase (CKB) are tools to organize and display knowledge in a readily accessible format; however, curators face the same challenges in comprehensively identifying clinically relevant information for curation. Natural language processing (NLP) has emerged as a promising direction for accelerating manual curation, but prior applications were often conceived as stand-alone efforts to automate curation, and the scope is often limited to simple entity and relation extraction. In this paper, we study the alternative paradigm of assisted curation and identify key desiderata to scale up knowledge curation with human-computer symbiosis.

Methods

We chose precision oncology for a case study and introduced self-supervised machine reading, which can automatically generate noisy training examples from unlabeled text. We developed a curation user interface (UI) for precision oncology and through iterative “curathons” (curation hackathons), conducted retrospective and prospective user studies for head-to-head comparison between manual and machine-assisted curation.

Results

Contrary to the prevailing assumption, we showed that high recall is more important for end-to-end assisted curation. In extensive user studies, we showed that assisted curation can double the curation speed and increase the number of findings by an order of magnitude for previously scarcely curated drugs.

Conclusion

We demonstrated that an iterative and thoughtful collaboration between professional curators and NLP researchers can facilitate rapid advances in assisted curation for precision medicine. Human-machine reading symbiosis can potentially be applicable to clinical care and research scenarios where curation is a major bottleneck.

Introduction

A perennial challenge to clinical researchers and practitioners is how to stay abreast with the rapid growth of biomedical research literature. This problem is particularly acute in the era of precision medicine, with an explosion of knowledge about the myriad molecular biomarkers and treatments. Manual curation requires domain expertise and is expensive and time-consuming. Natural language processing (NLP) has emerged as a promising direction for accelerating curation. However, prior work tends to consider NLP as a stand-alone approach for automatic curation. The scope is often limited to entity search or simple relation extraction (e.g., binary relations within single sentences), focusing on extracting common findings with high precision^{1,2}.

In this paper, we study the alternative paradigm of assisted curation, where curation is approached as an interactive collaboration between human experts and NLP-based machine readers. Machine readers can scan millions of documents in seconds, reducing the giant haystack to a small set of candidate needles. Such candidate facts are presented in a curation user interface (UI), which human experts can quickly verify by essentially clicking a button. The curation decisions provide feedback to further improve machine reading accuracy.

The emphasis on human-computer symbiosis has prompted several significant design changes. Attaining high coverage (i.e., high sensitivity/recall) is of paramount importance in machine reading, as we are striving to leave no facts behind. Precision (i.e., positive predictive value) still matters for curation efficiency, but even modest precision such as 50% can lead to substantial efficiency gain, while ensuring excellent coverage in the long tail of scientific findings. This opens up opportunities to leverage

the latest progress in NLP and machine learning, such as self-supervised machine reading, which uses prior knowledge to automatically generate noisy training examples from unlabeled text.

We conduct a case study on precision oncology and show how an iterative and thoughtful collaboration by professional curators and NLP researchers has led to rapid advances in assisted curation for molecular tumor boards and precision medicine.

Motivated by end-to-end curation needs, we expanded the scope of self-supervised machine reading to cross-sentence, document-level extraction of complex n-ary relations, compared to prior work that often focuses on extracting entities or simple relations³. Without any annotated examples, our self-supervised machine reader already attains high recall and reasonable precision. We developed a curation UI for precision oncology through iterative “curathons” (curation hackathons). In extensive user studies, we show that assisted curation can double the curation speed in general and increase the number of findings by an order of magnitude for previously scarcely curated drugs.

Methods

Precision oncology has seen an explosion of new findings on how tumors with a certain genomic biomarker (e.g., a point mutation or fusion) might respond to a given treatment. Manual curation of such knowledge typically starts with an explicit triaging step, e.g., by querying the PubMed search engine with relevant keywords such as “cancer”, “mutation” and a drug name. Each query might return hundreds of PubMed papers for a given drug and curation of each paper can take hours.

In assisted curation, we instead apply a machine reader to scan all PubMed articles and identify candidate findings, which curators can verify in a user interface (Figure 2). Each relation candidate is shown in context, with relevant entities highlighted. Users can also check the full article for more information. Curators can simply click a button to accept or reject the proposed extraction, thus greatly accelerating curation. The machine reading system and the assisted curation UI both stem from Project Hanover in Microsoft Biomedical NLP, and we refer to them as Hanover in this paper.

Self-supervised machine reading

The machine reader is powered by an NLP system that learns to extract such findings from text. Standard information extraction focuses on entity search or binary relations within single sentences. In precision oncology, as in many other biomedical scenarios, we need to extract complex n-ary relations that may span multiple sentences, such as the aforementioned drug-gene-mutation relation, which can be framed as a classification problem (whether a given text span mentions drug response given tumor gene mutation for a drug-gene-mutation tuple). Standard supervised learning requires manually annotating a large number of input-output examples. Instead, we apply self-supervised machine reading, which leverages an available knowledge base (KB) of known drug-gene-mutation relations to automatically annotate noisy training examples in unlabeled text. For any high-value relation, manually curated KBs generally exist. While manual curation is hard to sustain and can produce sparse coverage, these KBs can be used as a seed to bootstrap machine reading. Such a self-supervised approach has proven effective in our prior work⁴, and we use the same method in this paper.

Specifically, we use three public KBs (CIVIC⁵, OncoKB⁶, GDKD⁷) for self-supervision, which contain 1,256 unique drug-gene-mutation relation tuples. For unlabeled text, we use the PubMed Central Open Access Dataset, accessed in June 2018, with 1.65 million full-text articles. To identify drug and gene mentions, we use a curated list of canonical names and aliases based on publicly available ontologies. For mutations, we use simple regular expressions, focusing on missense mutations in this study. For training, we identify drug-gene-mutation tuples that appear nearby in text and automatically label an instance as positive if the tuple is in the KBs and negative otherwise. We focus on co-occurring mentions within a paragraph and leave cross-paragraph extraction to future work. We subsample the negative instances to ensure a balanced training set. This yields 4,906 positive and 13,371 negative examples, respectively, which we use to train a recurrent neural network (long short-term memory or LSTM).

To support end scenarios such as molecular tumor boards, curators might need more than drug-gene-mutation relations, such as the evidence type and strength (e.g., whether the finding stems from a human trial, the sample size, etc.). An automated curation system would have to tackle all these required aspects at once. In assisted curation, however, it suffices to focus on the core drug-gene-mutation relation initially, as curators can easily check the additional aspects during verification. This can already attain the majority of efficiency gain in curation, as most irrelevant papers do not contain any mention of the core relation.

User study

We applied the self-supervised machine reader to PubMed-scale knowledge extraction and assessed extraction quality and potential efficiency gain from assisted curation. For accuracy estimate, we sampled the machine-reading results and conducted several curathons with professional curators of the JAX Clinical Knowledgebase (CKB)⁸ to annotate whether a candidate drug-gene-mutation relation instance was “Curation Worthy”, “Entails, but not for curation”, or “Not Entail”. A candidate relation is considered “Curation Worthy” when it is described in the text (i.e., machine reading is correct), and it is deemed valid for curation into the database by the curators. “Entails, but not for curation” indicates that while the candidate relation is indeed described in the text, it is deemed not fit for curation given additional context in the paper (e.g., insufficient evidence strength). “Not Entails” indicates the candidate relation is not described in the text (i.e., machine reading is wrong).

For coverage estimate, we identified unique drug-gene-mutation ternary relations in machine reading results and compared them with those in CKB. For relations not in CKB, we used the above sample precision to estimate the curation-worthy portion. We conducted a series of controlled user studies to estimate potential efficiency gain from assisted curation. To ensure fair comparison, two curators with comparable curation efficiency and familiarity with both curation methods participated in the study. Both manual curation and assisted curation focus on missense mutations and PubMed and PMC text published in or after 2014.

Results

In each curathon and user study, we exclude relations that are already curated in CKB. In total, CKB curators have annotated 1147 candidate drug-gene-mutation relation

instances from self-supervised machine reading (Hanover). 749 are considered entailed, among which 351 are deemed curation-worthy. Consequently, we estimate that machine reading attains a precision of 65.3% and a curation-worthy proportion of 30.6%. Generally, CKB curators only need 1-2 minutes to validate each instance using the assisted curation UI (Figure 2).

Overall, we have applied machine reading to all publicly available PubMed and PMC papers, including 32.4 million abstracts and 3,671,360 full-text articles. This yields 27,308 unique ternary drug-gene-mutation relations, with 2,878 already appearing in CKB and 7,329 among the remainder estimated to be curation-worthy. Figure 3 compares curation-worthy coverage by CKB manual curation and Hanover machine reading from publicly available PubMed papers over the years. In total, Hanover extracts an estimated 10,207 unique curation-worthy relations, compared to 4,493 in CKB curated from the same set of PubMed text, with an estimated addition of 7,329 new curation-worthy relations, or 163% increase in coverage.

Finally, we conduct a series of controlled user studies to compare the efficiency of manual and assisted curation. We select 27 drugs that are relatively new and less well curated. In each user study, we always allotted equal time for manual and assisted curation. In the first two studies, the curators can choose the drugs to work on freely. In the final user study, we also pre-specify the allotted time for each drug in both manual and assisted curation, to enable more head-to-head comparison. The allotted time is set based on estimated abundance of findings. We set a minimum of 30 minutes for each drug, as manual curation requires significant time to conduct the initial triaging step. Assisted curation does not have such overhead and for a few drugs curators finished

validating all candidate instances ahead of the allotted time. Table 1 shows the detailed results. Assisted curation doubles the speed, yielding a total of 351 unique findings, compared to 180 in manual curation. It also uses less time, 669 vs 700 minutes.

As expected, manual curation covers far fewer PubMed articles compared to assisted curation, 78 compared to 221 articles. While assisted curation already doubles the speed, there is much room for further improvement. Often, manual curation benefits from curating a large number of relations from a single paper (e.g., 14 ternary relations from Table 9 in PMID 22734072 for drug tretinoin). Our current assisted curation UI, however, organizes candidate relations in the order of drug, gene, mutation. Therefore, findings from one paper might be scattered around based on the associated drugs and genes. A paper-centric view might enable much faster assisted curation.

Moreover, as mentioned before, while our machine reading precision is quite high (65.3%), the curation-worthy proportion is substantially lower (30.6%). This is because the current scope focuses on the core drug-gene-mutation relation, which does not account for additional aspects such as evidence type and strength. This suggests additional opportunities to improve assisted curation efficiency by expanding the scope of machine reading.

Discussion

Our study demonstrates the promise of machine-reading assisted curation in accelerating knowledge curation as in precision oncology for molecular tumor boards. Machine reading substantially increases curation coverage, and curators no longer need to undergo a repetitive and time-consuming process to scan a large swath of mostly irrelevant text. Instead, they can start with candidate relations with provenance

and quickly validate them in a curation-friendly UI, which can already double the curation speed. There are many opportunities to further increase the efficiency gain. Prior work in biomedical knowledge extraction, including commercially available tools, often focuses on entity search or simple relations such as binary drug-gene relations. For complex knowledge curation, as in precision oncology, they are typically insufficient to provide a significant boost in curation efficiency. For example, for one example drug, an entity search approach will yield over 800 PubMed articles that mention the given drug, as well as some gene mutations in the cancer context. Curating them all would have required over a thousand expert hours. By contrast, our machine reading system identifies 38 candidates, which curators finish validating in about an hour and identify 21 curation-worthy findings, whereas previously CKB only contained two findings for the given drug.

There are several limitations in our study. First, our text corpus is limited to publicly available PubMed text. There is much more biomedical text behind the paywall and applying machine-reading assisted curation to them can potentially extract even more knowledge. For simplicity, we focus on point mutations in this study, but in preliminary results we find that machine reading is also effective in extracting complex variants, such as fusion and indel. We focus on the core drug-gene-mutation relations in the initial exploration, but machine reading can be further expanded to include other aspects that are important for ascertaining curation-worthiness.

Reference

1. Lever J, Jones MR, Danos AM, et al: Text-mining clinically relevant cancer biomarkers for curation into the CIViC database. *Genome Med* 11:78, 2019
2. Lever J, Zhao EY, Grewal J, et al: CancerMine: a literature-mined resource for drivers, oncogenes and tumor suppressors in cancer. *Nat Methods* 16: 505–507, 2019
3. Wei C-H, Harris BR, Li D, et al: Accelerating literature curation with text-mining tools: a case study of using PubTator to curate genes in PubMed abstracts. *Database* 2012:bas041, 2012
4. Jia R, Wong C, Poon H: Document-level N-ary relation extraction with multiscale representation learning. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Volume 1, 10.18653/v1/n19-1370
5. Griffith M, Spies NC, Krysiak K, et al: CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. *Nat Genet* 49:170-174, 2017
6. Chakravarty D, Gao J, Phillips SM, et al: OncoKB: A Precision Oncology Knowledge Base. *JCO Precis Oncol* 2017:PO.17.00011, 2017
7. Dienstmann R, Jang IS, Bot B, et al: Database of genomic biomarkers for cancer drugs and clinical targetability in solid tumors. *Cancer Discov* 5:118-123, 2015
8. Patterson SE, Liu R, Statz CM, et al: The clinical trial landscape in oncology and connectivity of somatic mutational profiles to targeted therapies. *Hum Genomics* 10:4, 2016

Figure Legends

Figure 1: Assisted curation applies machine reading to reduce a giant haystack of source documents to a few candidate findings, which human experts can quickly validate through a curation user interface.

Figure 2: Assisted curation user interface for precision oncology. Users can verify a candidate finding (how a tumor with a given gene mutation might respond to a drug) by clicking a button.

Figure 3: Comparison of coverage by Clinical Knowledgebase (CKB) and self-supervised machine reading (Hanover), as measured by unique drug-gene-mutation relations curated from PubMed articles over the years. If a finding appears in multiple publications, the first will be counted. The apparent dip in 2020 is due to embargo (delay in PMC release of full-text articles).

Figure 4: Comparison of curation coverage from publicly available PubMed articles, in unique drug-gene-mutation relations (machine reading with estimated curation-worthy results).

Table 1: Comparison of manual and assisted curation efficiency in controlled user studies. In the first two studies, curators were given equal time (8 hours) for manual and assisted curation, respectively. In the final user study, we also pre-specified the allotted time for each drug to make the comparison more head-to-head.

Figure 1

Figure 2

Project Hanover

User microsoft

Logout

Evidence Type

☐ Clinical
 ☐ Preclinical Patient Derived Xenograft
 ☐ Preclinical Patient Derived Cell Culture
 ☐ Preclinical Cell Line Xenograft
 ☐ Preclinical Cell Line Culture
 ☐ Unknown

Publication Type

☒ Primary
 ☐ Review

Table/Text Type

☒ Table
 ☒ Text

Drugs (715)

10058-44

2-methoxyestradiol

5-fluoropyrimidine

79-6

a-1155463

a-1210477

a-395

a-485

abbv-075

abemaciclib

abiraterone

abi001

abt-263

abt-348

abt-737

ac-93253 iodide

talazoparib

Genes (13)

ATR

CTNNB1

IDH1

JAK2

KMT1B

KRAS

MAP2K1

MPL

PARP1

RAD51C

STAG2

TBCE

TP53

Variant

PubMed ID

Score

Level of evidence

a146t

28566428

0.78

To explore whether MEK could re-sensitize PARPi resistant cells to effects of PARPi, we developed PARPi resistant cells by culturing highly PARPi sensitive cells (UWB1.289 and A27980CP, both RAS wild type, see Fig. 2) in the continued presence of BMN673 for 3 to 4 months, at which time drug resistant clones emerged. A2780CP PARPi resistant (A2780CP_R) and UWB1.289 PARPi resistant (UWB1.289_R) clones were highly resistant to BMN673 and cross resistant to olaparib (Fig. 3A-B). RPPA analysis demonstrated that RAS / MAPK pathway activity (increased pMEK, pBAD, and pFOXO3a (inactive form)) was upregulated in PARPi resistant clones (Fig. 3C). Moreover, resistant clones showed lower total FOXO3a and BIM, as expected from increased RAS / MAPK pathway activity. The decreased PAR and PARP1 expression in the resistant cells could also contribute to PARPi resistance, as PARP1 expression is associated with PARPi sensitivity (22). Western blotting confirmed increased RAS / MEK pathway activity with concomitant decreases in FOXO3a and BIM in resistant cells (Fig. 3D). Overall, the signaling changes in long-term PARPi resistant cells exhibited many similarities to adaptive responses to short-term PARPi treatment (see Fig. 1). Despite increased RAS / MEK pathway activity, KRAS sequencing demonstrated that the resistant lines did not acquire classical activating KRAS mutations. However, deep NGS sequencing as well as Sanger sequencing of individual PARPi resistant clones from A2780CP_R demonstrated the presence of KRAS A59T and MAP2K1 A283T in 19, 11 and 6 % of cells respectively but not in A2780CP parental cells. Importantly, prolonged culture of the lines without PARPi resulted in loss of the mutant KRAS and MAP2K1 clones. The KRAS A546T mutant has been demonstrated to be modestly activating (30). The selection of KRAS mutations in a PARPi resistant line supports the concept that RAS mutations and RAS / MAPK pathway activation is a key mediator of PARP resistance. As expected by increases in RAS / MAPK activity in PARPi resistant cell lines and KRAS and MAPK1 mutations, A2780CP_R were markedly more sensitive and UWB1.289_R were modestly more sensitive to MEK (Fig. 3E-F). MEK re-sensitized both PARPi resistant clones to PARPi (Fig. 3E-F). Thus

Duration Worthy

Entails, but not for Duration

Not Entails

Clear

This paragraph describes an observed relation between talazoparib, KRAS and a146t

Add notes

Updated by microsoft 24 days ago (Jul 25, 2019 3:48:29 PM)

a59t

28566428

0.56

To explore whether MEK could re-sensitize PARPi resistant cells to effects of PARPi, we developed PARPi resistant cells by culturing highly PARPi sensitive cells (UWB1.289 and A27980CP, both RAS wild type, see Fig. 2) in the continued presence of BMN673 for 3 to 4 months, at which time drug resistant clones emerged. A2780CP PARPi resistant (A2780CP_R) and UWB1.289 PARPi resistant (UWB1.289_R) clones were highly resistant to BMN673 and cross resistant to olaparib (Fig. 3A-B). RPPA analysis demonstrated that RAS / MAPK pathway

Contact Us | Privacy & cookies | © Microsoft 2019

Figure 3

Figure 4

Table 1

Drugs	Allotted Time (min)	Manual Yield	Assisted Curation Yield	Machine Reading Precision (%)	Machine Reading Curation-Worthy Proportion (%)
User Study 1					
anastrozole	-	7	7	75.0	25.0
axitinib	-	18	27	84.6	51.9
bafetinib	-	1	21	97.4	55.3
bevacizumab	-	8	2	27.6	6.9
alectinib	-	2	12	50.0	33.3
alvespimycin	-	0	33	70.7	56.9
bortezomib	-	0	32	56.9	31.4
exemestane	-	1	-	-	-
bosutinib	-	7	-	-	-
User Study 2					
tretinoin	-	38	28	61.6	32.6
vorinostat	-	5	37	63.0	37.0
tanespimycin	-	10	21	81.0	33.3
trastuzumab emtansine	-	4	0	60.0	0.0
rociletinib	-	2	3	94.1	17.7

valproic acid	-	3	6	65.3	12.2
Sp600125	-	4	-	-	-
User Study 3					
sp600125	90	3	1	13.9	2.8
infigratinib	30	2	7	80.0	46.7
exemestane	30	3	8	100.0	21.6
rituximab	45	0	10	56.7	33.3
epigallocatechin gallate	45	2	5	79.1	11.6
bosutinib	60	2	4	63.5	7.7
pazopanib	60	1	18	56.3	37.5
erlotinib	90	11	4	48.0	5.3
crenolanib	30	8	16	86.4	72.7
bortezomib	60	8	8	51.4	21.6
icotinib	30	6	6	85.0	30.0
quizartinib	45	17	9	52.4	42.9
genistein	30	1	0	71.4	0.0
navitoclax	45	3	17	80.0	68.0
tanespimycin	30	3	9	82.4	52.9
		180	351		