

DC³ – A Diagnostic Case Challenge Collection for Clinical Decision Support

Carsten Eickhoff
Brown University
codiag AG
USA

Floran Gmehlin
Anu V. Patel
Jocelyn Boullier
codiag AG
Switzerland

Hamish Fraser
Brown University
USA

ABSTRACT

In clinical care, obtaining a correct diagnosis is the first step towards successful treatment and, ultimately, recovery. Depending on the complexity of the case, the diagnostic phase can be lengthy and ridden with errors and delays. Such errors have a high likelihood to cause patients severe harm or even lead to their death and are estimated to cost the U.S. healthcare system several hundred billion dollars each year.

To avoid diagnostic errors, physicians increasingly rely on diagnostic decision support systems drawing from heuristics, historic cases, textbooks, clinical guidelines and scholarly biomedical literature. The evaluation of such systems, however, is often conducted in an ad-hoc fashion, using non-transparent methodology, and proprietary data.

This paper presents DC³, a collection of 31 extremely difficult diagnostic case challenges, manually compiled and solved by clinical experts. For each case, we present a number of temporally ordered physician-generated observations alongside the eventually confirmed true diagnosis. We additionally provide inferred dense relevance judgments for these cases among the PubMed collection of 27 million scholarly biomedical articles.

KEYWORDS

Diagnostics, Rare Diseases, Corpus, Dataset

ACM Reference Format:

Carsten Eickhoff, Florian Gmehlin, Anu V. Patel, Jocelyn Boullier, and Hamish Fraser. 2019. DC³ – A Diagnostic Case Challenge Collection for Clinical Decision Support. In *The 2019 ACM SIGIR International Conference on the Theory of Information Retrieval (ICTIR '19), October 2–5, 2019, Santa Clara, CA, USA*. ACM, New York, NY, USA, 4 pages. <https://doi.org/10.1145/3341981.3344239>

1 INTRODUCTION

Diagnostic errors refer to the failure to establish an accurate and timely explanation of the patient's health problem(s) or to communicate that explanation to the patient [13]. With relative shares

of up to 40%, several studies report diagnostic errors to constitute the largest and most impactful source of avoidable primary care error [15, 16, 21], leading to the most considerable share of claims [3, 18, 20] against primary care physicians. In the USA, an annual economic damage of hundreds of billions (a sizable portion of the country's overall health spendings) is attributed to diagnostic errors [11]. Given the key importance of correctness and timeliness of primary care diagnosis, a large body of work has been investigating systematic reasons for misdiagnoses in this setting. Numerous studies list low disease prevalence as one of the top causes of diagnostic errors and delays in primary care [8, 17, 20], as uncommon diagnoses may be overshadowed by more prevalent ones in the cognitive diagnostic process [7, 17]. In an effort to improve patient safety, there are frequent calls for more effective diagnostic processes in primary care [10, 27], involving a greater utilization of electronic health record (EHR) and clinical decision support systems [14]. Complex patients with non-specific presentations, multiple co-morbidities, and rare conditions are assumed to be at an especially high risk of receiving a delayed or inaccurate diagnosis [13].

Clinical decision support systems aim to help health professionals in addressing particularly challenging diagnosis or treatment needs [2, 4, 12, 26]. In order to assess the accuracy of such systems, annotated examples of patient case information are required. To date, there are only few such resources available to researchers. As a consequence, many clinical decision support system evaluation campaigns rely on small, outdated or proprietary sources of data, making their findings difficult to verify and reproduce.

This paper presents DC³, a collection of 31 extremely difficult diagnostic case challenges, that were manually compiled and solved by clinical experts. For each case, there are a number of temporally ordered physician-generated observations alongside the eventually confirmed true diagnosis. We additionally provide inferred dense relevance judgments for these cases among the PubMed collection of 27 million scholarly biomedical articles.

2 COMPARISON TO EXISTING COLLECTIONS

The vast majority of clinical decision support systems are trained and evaluated on proprietary samples of patient data that, for reasons of confidentiality, cannot be released to the research community. There are, however, a number of openly available collections that deserve mentioning.

TREC CDS. The TREC Clinical Decision Support (CDS) track [19] was run from 2014 to 2016 and tasked participating systems to retrieve biomedical literature in response to 90 natural-language

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

ICTIR '19, October 2–5, 2019, Santa Clara, CA, USA

© 2019 Association for Computing Machinery.

ACM ISBN 978-1-4503-6881-0/19/10...\$15.00

<https://doi.org/10.1145/3341981.3344239>

patient descriptions. The patients stem from general intensive care populations with relatively common indications. For the majority of these patients no explicit target diagnosis information is available. All patient descriptions are single-shot narratives in the style of an anamnesis without temporally ordered findings.

CLEF eHealth. Similar to the TREC CDS track, the 2013 to 2015 editions of CLEF’s eHealth challenge [23] offered a patient-centric information retrieval task. Given a minimal patient profile, the participants were tasked with retrieving relevant literature related to the case.

MIMIC-III. The MIMIC critical care database [6] contains electronic records of 59,000 intensive care admissions collected between the years of 2001 and 2012. Each admission specifies a list of diagnostic and procedural ICD codes that were designed for billing purposes but can be used as targets for diagnostic decision support. Given the intensive care domain, the range of observed diagnoses is limited in comparison to the complex cases outlined here.

i2b2. The i2b2 initiative has been launching a number of classification and information extraction benchmarking campaigns for which annotated corpora of de-identified patient records were provided. Most studied tasks focus on extracting patient properties such as smoker status [25] or obesity [24] that are of only limited interest in the diagnostic decision support use case. Others are limited to a single class of diagnoses such as heart disease [22] and do not lend themselves to wide-coverage primary care research.

To the best of our knowledge, at the time of writing this paper, there exists no publicly available dataset of challenging authentic diagnostic episodes that additionally provide both confirmed diagnoses as well as dense query-document relevance judgments for a realistically-sized document collection. DC³ aims to close this gap.

3 DATASET DETAILS

The dataset compiles 31 especially challenging diagnostic cases encountered at Massachusetts General Hospital (MGH) in Boston, MA between the years 2013 and 2018.

3.1 Cases

Each case is described in a number of topically coherent paragraphs that correspond to the sections usually found in case notes. For the purpose of data extraction, we assume each paragraph to denote an episode in a health record/note entry. Example paragraphs include presenting complaint, history of presenting complaint, examination or investigation and may have been authored by changing physicians (*e.g.*, the first note describes the patient’s anamnesis taken by the emergency department’s staff while the following note might be written by a radiologist, discussing the findings of a CT scan, *etc.*). The cases featured in this corpus are all complex and difficult. While a good proportion of the featured diagnoses are rather uncommon in developed-world hospitals such as MGH (*e.g.*, 16 cases of infectious diseases, or a lead poisoning), they can be more frequently observed in large parts of the developing world. Aside from low-prevalence, many cases have multiple correct diagnoses that jointly account for the patient’s symptoms. We represent diagnoses in terms of the corresponding concept unique identifiers (CUI) of the Unified Medical Language System (UMLS). The average case in the collection has 6.9 target CUIs. Finally, we annotate the 2018

snapshot of the National Library of Medicine’s PubMed database, identifying all papers whose abstracts mention any of the target diagnoses. In order to achieve this, we rely on a proprietary medical named entity recognition system. Any paper whose title or abstract mention the target diagnoses of a case is considered relevant for that case. The assumption here is that presenting a physician with a paper mentioning the correct diagnosis for the current patient will bring that diagnosis to their attention and thereby increase the chance of them testing for and eventually confirming it. Unlike many existing IR benchmarking collections, that rely on pooled manual relevance assessment, this approach results in dense¹ relevance labels. The median number of relevant documents per case is 3,597. Especially for outlier cases with many target diagnoses, such numbers are much higher than what is observed in typical Web search collections with only partial relevance labels. Table 1 gives a complete overview of all cases. We did not perform any manual sub-selection of cases and, instead, included all currently published MGH case challenges.

The narrative content of the case notes is written by the treating physicians of the original cases and is composed to reflect the temporal order of discoveries, hypotheses and tests performed, but does not directly reveal the target diagnosis.

3.2 Distribution Format

All cases presented in DC³ are originally published in the New England Journal of Medicine’s *Case Challenge* section. The copyright remains with them and we do not redistribute any of the case content directly. Instead, we provide the research community with a convenient Python script² that downloads the publicly available case challenges and organizes them in the form of a JSON file. Figure 1 shows an example of the resulting format. Additionally, we collected inferred dense relevance judgments for the 2018 snapshot of the National Library of Medicine’s PubMed database of 27 million scholarly biomedical articles. These relevance judgments are provided in standard `trec_eval` format.

4 TASKS & BASELINE PERFORMANCE

In this section, we will describe a range of possible experiments on the DC³ collection. We begin with a patient-centric information retrieval task [1, 9] in which we measure the ranking performance of models that take the case description as a query and retrieve scholarly literature articles conducive to making the correct diagnosis. Afterwards, we cast the diagnostic decision support task as a supervised text classification problem in which we model the posterior probability of observing the case description given a diagnosis-specific classifier.

4.1 Patient-centric Document Retrieval

This task is similar to the one studied in the TREC Clinical Decision Support (CDS) track [19]. We use Lucene to index all 27M PubMed abstracts and use the full case description as a query. Table 2 reports

¹There is the possibility of NER false negatives that would lead to missed potentially relevant documents. Given the generally high performance of this system, we consider this a minor risk to corpus quality.

²<https://github.com/codiag-public/dc3>

Table 1: DC³ Case Details

Case ID	# Notes	Final Diagnosis	# CUIs	# rel. Docs.
1	10	Histoplasma capsulatum infection	3	121
2	11	Necrotizing lymphadenitis, with features consistent with histiocytic necrotizing lymphadenitis (Kikuchi-Fujimoto disease) and scattered EBV-positive cells	3	3,918
3	7	Infective endocarditis and infectious aortitis due to Staphylococcus aureus	18	10,767
4	8	Lead poisoning	1	2,387
5	13	Measles	7	13,466
6	7	Wilson's Disease	1	1,298
7	11	Lemierre's syndrome caused by Fusobacterium necrophorum, with cavernous-sinus thrombophlebitis, carotid-artery thromboarteritis, and abscesses of the parotid gland and subperiosteal orbit	9	276
8	7	Acute anaphylaxis due to a hepatic hydatid cyst caused by Echinococcus granulosus.	10	11,638
9	6	Invasive Neisseria meningitidis infection and primary C8 deficiency	5	3,149
10	10	Perforation of the right ventricular wall (by an implantable cardioverter-defibrillator lead)	2	319
11	8	Borrelia miyamotoi infection and possible Borrelia burgdorferi infection	4	11,044
12	11	Disseminated pulmonary blastomycosis involving the hilar lymph nodes and spleen, early hepatic cirrhosis, and acute tubular necrosis	5	55,750
13	10	Disseminated Mycobacterium bovis infection	6	54,359
14	14	Chronic recurrent abdominal pain caused by intermittent torsion of an accessory spleen	4	794
15	25	Tuberculous enteritis	4	6,501
16	7	Mixed-cellularity subtype of classic Hodgkin's lymphoma and Epstein-Barr virus infection	10	7,538
17	9	Secondary syphilis with neurologic, ocular, and otologic involvement	13	1,774
18	8	Milk Alkali Syndrome	1	343
19	9	Acute HEV infection (Acute Viral Hepatitis)	13	9,222
20	12	Granulomatous amebic encephalitis, caused by acanthamoeba species. Sarcoidosis (old, burned out) involving heart, lungs, and spleen. Coronary arteriosclerosis with stent stenosis. Papillary renal-cell carcinoma and benign biliary hamartoma	29	219,540
21	8	Primary adrenal insufficiency (Addison's disease)	2	864
22	7	IgA vasculitis (Henoch Schonlein Purpura - HSP)	2	23,123
23	11	Acute Leptospirosis	7	3,597
24	7	Acute and chronic cholecystitis and extensive cholelithiasis with transmural gallbladder inflammation	19	19,907
25	11	Combined inherited and acquired thrombotic thrombocytopenic purpura	4	3,233
26	10	Congenital rubella syndrome	6	6,545
27	8	Mycobacterial epididymo-orchitis due to Mycobacterium tuberculosis	10	91,855
28	6	Digoxin toxicity	2	387
29	8	Well-differentiated pancreatic neuroendocrine tumor, grade 1	3	3,420
30	6	Complete androgen insensitivity syndrome	5	1,219
31	11	Lyme meningoradiculitis	2	135

ranking performance of a range of well-known retrieval models in terms of nDCG [5] scores at this task.

4.2 Classification

In an alternative take on the diagnostic decision support problem, we train a range of disease-specific classifiers on the basis of Pubmed abstracts. We collect all Pubmed abstracts mentioning the target diagnosis as training data and assign the class of maximum posterior

probability. This classification problem becomes more difficult as increased numbers k of target diagnoses are being considered. We include the top $k = \{500, 1000, 2000\}$ most frequent diagnoses as observed in Pubmed and compare the performance of Naïve Bayes, Logistic Regression and Support Vector Machine classifiers. This selection of classification methods is by no means exhaustive and many more modern and sophisticated techniques are expected to perform better. This overview merely demonstrates the complexity

```

1 {"version": "1.0",
2   "download-date": "2018-12-08",
3   "cases": [
4     {"case-id": 1,
5      "title": "A 29-Year-Old Man with...",
6      "diagnosis": "Histoplasma capsulatum infection",
7      "CUIs": ["C0153261", "C0153262", "C0153268"],
8      "notes": [
9        {"note-id": 1,
10         "content": "(Medicine): A 29-year-old..."},
11        {"note-id": 2,
12         "content": "The patient was reportedly..."},
13        ...

```

Figure 1: An example of the DC³ case file format.

Table 2: Patient-centric literature retrieval results.

Model	nDCG
TF-IDF	0.37
LambdaMART	0.41
DRMM	0.42

Table 3: Classification results for different choices of k .

Model	500	1000	2000
Naïve Bayes	0.13	0.08	0.04
Logistic Regression	0.14	0.08	0.06
SVM	0.17	0.09	0.07

of the task and the demand for innovation in order to truly support unconstrained primary care diagnosis. Table 3 reports the results of this comparison in terms of F_1 scores. In addition to performance differences between models, increased numbers k of considered target diagnoses significantly increase classification difficulty.

5 CONCLUSION

In this paper, we present the first version of DC³, a diagnostic case challenge collection for evaluation of clinical decision support systems. The corpus compiles 31 challenging cases from Massachusetts General Hospital in Boston, MA alongside their true underlying diagnoses. As an especially interesting property, we share inferred dense relevance judgments for these cases and the 2018 snapshot of the NLM’s PubMed database, allowing for robust and reproducible benchmarking of clinical decision support techniques. In an effort to gauge the collection’s difficulty, we investigated two common tasks of interest: Patient-centric literature retrieval, and supervised classification for clinical decision support.

This paper describes a piece of early work in progress that has several limitations to be addressed in the future. (1) Purely inferred relevance judgments do not replace manual expert annotations. They do however offer a powerful means of training retrieval systems that aim to bring the true diagnosis to the physician’s attention. (2) Concentrating exclusively on complex cases does not reflect the full spectrum of daily diagnostic tasks encountered by physicians but may help address those cases that doctors struggle with most. (3) The classification and retrieval methods presented in this collection paper are not meant to reflect competitive solutions to the problem but rather aim to illustrate task complexity.

REFERENCES

- [1] Prakrit Baruah, Riya Dulepet, Kyle Qian, and Carsten Eickhoff. Brown university at trec precision medicine 2018. In *Proceedings of the 27th Text Retrieval Conference (TREC)*, 2018.
- [2] Ferenc Galkó and Carsten Eickhoff. Biomedical question answering via weighted neural network passage retrieval. In *Proceedings of the 40th European Conference on Information Retrieval (ECIR)*, Springer, 2018.
- [3] Tejal K Gandhi, Allen Kachalia, Eric J Thomas, Ann Louise Puopolo, Catherine Yoon, Troyen A Brennan, and David M Studdert. Missed and delayed diagnoses in the ambulatory setting: a study of closed malpractice claims. *Annals of internal medicine*, 145(7):488–496, 2006.
- [4] Paulina Grnarova, Florian Schmidt, Stephanie L Hyland, and Carsten Eickhoff. Neural document embeddings for intensive care patient mortality prediction. *NIPS Workshop on Machine Learning for Health*, 2016.
- [5] Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM TOIS*, 20(4):422–446, 2002.
- [6] Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. MIMIC-iii, a freely accessible critical care database. *Scientific data*, 3:160035, 2016.
- [7] David A Jopp and Christopher B Keys. Diagnostic overshadowing reviewed and reconsidered. *American Journal on Mental Retardation*, 106(5):416–433, 2001.
- [8] Olga Kostopoulou, Brendan C Delaney, and Craig W Munro. Diagnostic difficulty and error in primary care? a systematic review. *Family practice*, 25(6), 2008.
- [9] Lorenz Kuhn and Carsten Eickhoff. Implicit negative feedback in clinical information retrieval. In *Proceedings of the ACM SIGIR Medical Information Retrieval Workshop*, 2016.
- [10] CY Lorincz, E Drazen, PE Sokol, KV Neerukonda, J Metzger, MC Toepf, L Maul, DC Classen, and MK Wynia. Research in ambulatory patient safety 2000–2010: a 10-year review. *Chicago, IL: American Medical Association*, 2011.
- [11] J Michael McGinnis, Leigh Stuckhardt, Robert Saunders, Mark Smith, et al. *Best care at lower cost: the path to continuously learning health care in America*. National Academies Press, 2013.
- [12] Alexander Meyer, Dina Zverinski, Boris Pfahringer, Jörg Kempfert, Titus Kuehne, Simon H Sündermann, Christof Stamm, Thomas Hofmann, Volkmar Falk, and Carsten Eickhoff. Machine learning for real-time prediction of complications in critical care: a retrospective study. *The Lancet Respiratory Medicine*, 6(12), 2018.
- [13] Engineering National Academies of Sciences, Medicine, et al. *Improving diagnosis in health care*. National Academies Press, 2016.
- [14] David E Newman-Toker and Peter J Pronovost. Diagnostic errors?the next frontier for patient safety. *Jama*, 301(10):1060–1062, 2009.
- [15] Robert L Phillips, Lori A Bartholomew, Susan M Dovey, GE Fryer, TJ Miyoshi, and LA Green. Learning from malpractice claims about negligent, adverse events in primary care in the united states. *BMJ Quality & Safety*, 13(2):121–126, 2004.
- [16] John Sandars and Aneez Esmail. The frequency and nature of medical error in primary care: understanding the diversity across studies. *Family practice*, 2003.
- [17] Urmimala Sarkar, Doug Bonacum, William Strull, Christiane Spitzmueller, Nancy Jin, Andrea López, Traber Davis Giardina, Ashley ND Meyer, and Hardeep Singh. Challenges of making a diagnosis in the outpatient setting: a multi-site survey of primary care physicians. *BMJ Qual Saf*, pages bmjqs–2011, 2012.
- [18] N Silk. What went wrong in 1,000 negligence claims. *Health care risk report*, 7:13–15, 2000.
- [19] Matthew S Simpson, Ellen M Voorhees, and William Hersh. Overview of the trec 2014 clinical decision support track. Technical report, DTIC Document, 2014.
- [20] Hardeep Singh, Traber Davis Giardina, Ashley ND Meyer, Samuel N Forjuoh, Michael D Reis, and Eric J Thomas. Types and origins of diagnostic errors in primary care settings. *JAMA internal medicine*, 173(6):418–425, 2013.
- [21] Hardeep Singh and Mark Graber. Reducing diagnostic error through medical home-based primary care reform. *Jama*, 304(4):463–464, 2010.
- [22] Amber Stubbs, Christopher Kotfila, Hua Xu, and Özlem Uzuner. Identifying risk factors for heart disease over time: Overview of 2014 i2b2/uthealth shared task track 2. *Journal of biomedical informatics*, 58:S67–S77, 2015.
- [23] Hanna Suominen, Sanna Salanterä, Sumithra Velupillai, Wendy W Chapman, Guergana Savova, Noemie Elhadad, Sameer Pradhan, Brett R South, Danielle L Mowery, Gareth JF Jones, et al. Overview of the share/clef ehealth evaluation lab 2013. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 212–231. Springer, 2013.
- [24] Özlem Uzuner. Recognizing obesity and comorbidities in sparse data. *Journal of the American Medical Informatics Association*, 16(4):561–570, 2009.
- [25] Özlem Uzuner, Ira Goldstein, Yuan Luo, and Isaac Kohane. Identifying patient smoking status from medical discharge records. *Journal of the American Medical Informatics Association*, 15(1):14–24, 2008.
- [26] Xing Wei and Carsten Eickhoff. Embedding electronic health records for clinical information retrieval. *arXiv preprint arXiv:1811.05402*, 2018.
- [27] Matthew K Wynia and David C Classen. Improving ambulatory patient safety: learning from the last decade, moving ahead in the next. *Jama*, 306(22), 2011.