npj | Digital Medicine

BRIEF COMMUNICATION OPEN

Real-time clinician text feeds from electronic health records

James T. H. Teo[1,2,3]✉, Vlad Dinu[3], William Bernal[1], Phil Davidson[1], Vitaliy Oliynyk[2], Cormac Breen[2], Richard D. Barker[1] and Richard J. B. Dobson[3]

Analyses of search engine and social media feeds have been attempted for infectious disease outbreaks, but have been found to be susceptible to artefactual distortions from health scares or keyword spamming in social media or the public internet. We describe an approach using real-time aggregation of keywords and phrases of freetext from real-time clinician-generated documentation in electronic health records to produce a customisable real-time viral pneumonia signal providing up to 4 days warning for secondary care capacity planning. This low-cost approach is open-source, is locally customisable, is not dependent on any specific electronic health record system and can provide an ensemble of signals if deployed at multiple organisational scales.

npj Digital Medicine (2021)4:35 ; https://doi.org/10.1038/s41746-021-00406-7

Analyses of search engine and social media feeds have been attempted for infectious disease outbreaks[1], but have been susceptible to artefactual distortions from health scares or keyword spamming in social media or the public internet[2–4]. In contrast to the unfiltered unmoderated freetext of the public internet, freetext in electronic health records (EHR) are less tempestuous because of access management to healthcare-professionals-only; text in an EHR is therefore richer in saliency, lower in non-specific noise and less prone to distortions. However EHR's often have inconsistent data standardisation with structured data mixed with unstructured free text, as well as a hybrid of modern and legacy closed systems. More complex systems using machine learning of observations from structured data have promising accuracy[5] but would be heavily dependent on hospitals having highly structured and interoperable EHR's. Traditional data aggregation methods rely on gold-standard cases generated from reporting mechanisms like structured case report forms for local or national registries.

We describe an approach using real-time aggregation of keywords and key phrases from freetext in electronic health records to produce a real-time signal during the Covid pandemic. This open-source system takes text from structured and unstructured fields in near-real-time from clinician-generated documentation in the electronic health records and does not require health data to be standardised into any ontology. This real-time symptom-based unstructured text aggregation could provide earlier warning as it avoids laboratory processing delays and undersampling bias (significant confounders during the early pandemic period).

A query was defined producing an aggregated count of patient documents containing symptom keywords and phrases for symptomatic Covid: ("dry cough", "pyrexia", "fever", "dyspnoea", "anosmia", "pneumonia", "LRTI", "lung consolidation", "pleuritic pain" and associated synonyms, acronyms and poecilonyms) normalised against patient documents not containing these phrases or containing negations of the phrases (e.g. "no dry cough", "no anosmia"). This was used to generate an index of signal enriched for symptom clusters of symptomatic Covid which compares favourably to the gold-standard data of laboratory samples at King's College Hospital (KCH) and Guys and St Thomas' Hospital (GSTT) (Fig. 1).

Cross-correlation with laboratory samples from each hospital peaked at 4 days at KCH (cross-correlation $r = 0.783$) and GSTT at 0 days before (cross-correlation $r = 0.757$) (Supplementary Fig. 1A) with a good cross-correlation between sites (Supplementary Fig. 1B).

Signal density distribution was also performed to quantify the power of the signal at KCH and GSTT (Supplementary Fig. 2A). As the KCH signal was derived from a mix of sources (A&E admissions and Inpatient text), this was decomposed into two distinct signals as well (Supplementary Fig. 2B) with a higher signal strength in records which have heading prompts to the text boxes (e.g. history, investigations, management) at GSTT Inpatients and KCH A&E compared to records with complete unprompted free text (KCH Inpatients). The decomposed KCH signals had multiple sub-peaks (likely related to weekday–weekend cycles in A&E) and an accumulating correlation for inpatients (due to the persistence of affected patients in hospital) (Supplementary Fig. 2C). This indicates that analogous freetext signals between institutions should not be directly compared. The text signal was also strongly correlated with data on Covid cases using regional datasets for London hospital admissions (Supplementary Fig. 3) while capturing the effects of multiple waves. Of note, seasonal influenza epidemics over the winters from 2018–2020 were detectable due to many overlapping text features (Supplementary Fig. 4) indicating non-specificity for viral pneumonias.

Previous attempts at text-based epidemic forecasting have been largely focused on search engine or social media trends from the public internet to forecast influenza, and to measure such trends against structured public health databases at national level[6–10]. Other approaches for population-level forecasting have used structured manual submissions to national public health systems like the Centre for Disease Control's Influenza-Like Illness Reports[9]. Our approach does not attempt to forecast but acts as a noisy real-time barometer of local clinical data which is adequate for local operational use at low-cost.

While we show how this approach operates at the scale of single healthcare organisations, it can be scaled to a whole system level either by combining locally generated customised real-time signals from multiple organisations or by centralising the clinical data first before generating the real-time freetext signal from a 'bag of words'. However depending on the coverage of the local

[1]Kings College Hospital NHS Foundation Trust, London, United Kingdom. [2]Guys & St Thomas Hospital NHS Foundation Trust, London, United Kingdom. [3]Institute of Psychiatry, Psychology and Neuroscience, Kings College London, London, United Kingdom. ✉email: jamesteo@nhs.net

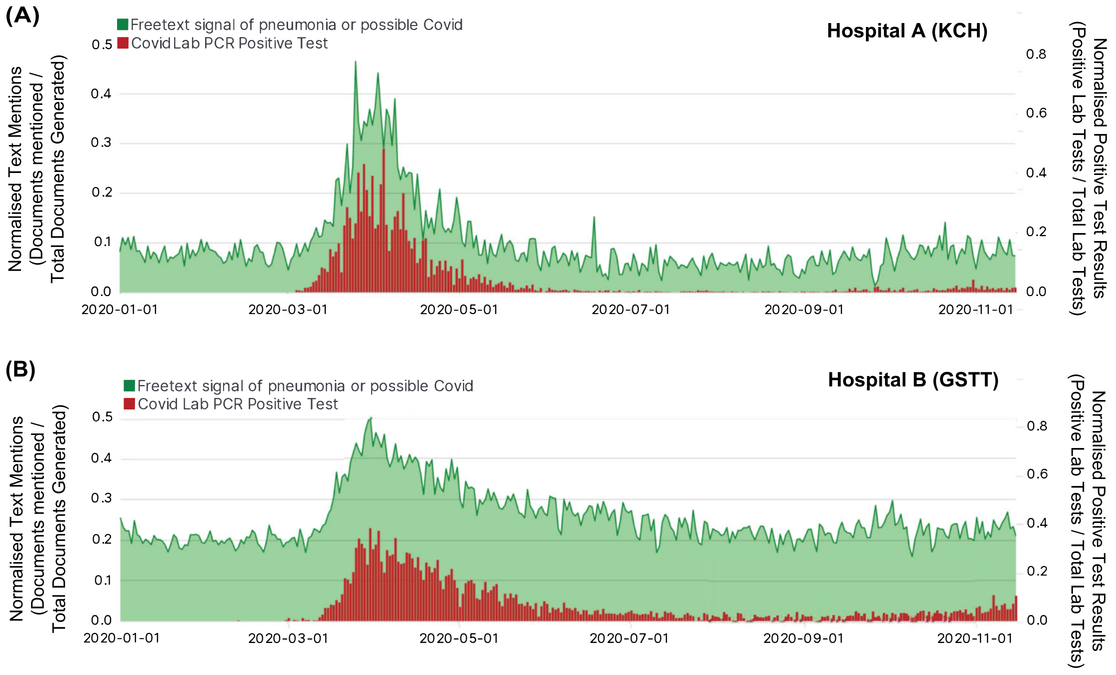

Fig. 1 Timeline of freetext signal for Covid-like symptoms from clinician text entries for two major London hospitals during the Covid pandemic. A Hospital A (Kings College Hospital, KCH) freetext signal (Green) and positive laboratory tests normalised against negative tests (Red) from 1st January 2020 till 12 November 2020; **B** Hospital B (Guys & St Thomas' Hospital, GSTT) freetext signal (Green) and positive laboratory tests normalised against negative tests (Red) from 1st January 2020 till 12 November 2020. KCH text signal was derived from both A&E and inpatients, while GSTT text signal was derived only from inpatients (A&E data not accessible). A figure with extension of the time period into January 2021 is available in Supplementary Fig. 6.

EHR over local care pathways, signals between hospitals may not be directly comparable. As such, we believe an ensemble of locally generated signals is superior as a single organisation is able to customise the locally derived signal for local operational purposes while simultaneously a wider health economy may derive utility from the ensemble of signals for other central planning purposes (mindful of the limits of comparing different organisations).

It is important to note limitations to our approach – firstly, ecological real-time freetext data is uncurated, and could be susceptible to distortions as certain phrases in freetext could produce artificial distortions (e.g. if ward names, job titles or operational pathways are named "Covid"). This risk is minimised in this study as the word "Covid" was not included from the signal in this study, but this could be tailored to the local context or dialect.

Secondly, one can only aggregate and harness knowledge that clinicians and patients consider of sufficient relevance or saliency to record; loss of smell and anosmia has been described in Covid[11]. Our text aggregator was able to detect an initial attenuated signal in the freetext in March 2020 without the benefit of recall bias (Supplementary Fig. 5). The converse is true, the signal was subsequently amplified by the media as subsequent scientific publication of the findings of anosmia on 17/05/2020 caused a second surge in the incidence of these phrases likely due to increasing clinician awareness.

This media-sensitive signal highlights that incident clinical language could become self-fulfilling–heightened awareness may increase the use of such words even in clinical text either through speculative differentials or checklists. We managed this by eliminating negation terms, combining multiple symptom phrases and focusing on formal documentation only (e.g. clinical episode summaries). Nonetheless an organisation should exercise caution against disseminating the exact words or phrases used to generate the signal to minimise 'hashtag meme'-like or 'trending' phenomenon amongst clinicians. Additionally, this means that the freetext signal would only be useful for evaluating prevalence during periods where there is no time-varying potentiation on clinician record-keeping by media or local policies.

Thirdly, this study as proof-of-concept was confined to secondary care settings; during the Covid first wave in March 2020, there were triage mechanisms to redirect low severity cases away from secondary care affecting visibility. For a global view of a health economy, text aggregators would also need to be deployed in primary care and emergency care providers.

In summary, we report a natural language approach of real-time clinical data that is flexible and scalable to feed dashboards of activity for capacity planning and can be generalised across organisations to provide early warning of future pandemic surges. This low-cost approach is open-source, is not fixed into any specific EHR system. It might be possible to deploy this at multiple organisations to provide an ensemble of signals.

METHODS

Platform

The text feed was generated by a locally deployed Cogstack instance[12] at Kings' College Hospital NHS Foundation Trust, UK consisting of an open-source toolkit for extracting text from EHR's (https://github.com/CogStack). A subsequent instance was deployed at Guys & St Thomas' Hospital NHS Foundation Trust, UK as validation in another hospital with different care pathways, different EHR and sociodemographics.

Text feed

A query was defined producing an aggregated count of patient documents containing symptom keywords and phrases for symptomatic Covid: ("dry cough", "pyrexia", "fever", "dyspnoea", "anosmia", "pneumonia", "LRTI", "lung consolidation", "pleuritic pain" and associated synonyms, acronyms and poecilonyms) normalised against patient documents not containing these phrases or containing negations of these phrases (e.g. "no dry cough"). This produces a signal which is enriched for symptom clusters of symptomatic Covid. Of note, affirmative diagnostic phrases (e.g. "confirmed Covid") were excluded from the query in this manuscript although these are used in practice to boost the signal. No machine learning was used to derive the query.

Presence of positive COVID PCR laboratory samples was used as the gold standard signal of cases. A similar query was constructed with local adaptations within a parallel Cogstack instance at GSTT.

Legal and ethical use of patient data

This project was conducted under public health requirements and operational service delivery of the hospitals as part of the UK's NHS (Control of Patient Information Regulations) 2002 (COPI) regulation to share confidential patient information to process under for COVID-19 purposes. Identifiable data was not evaluated. Additional guidance and support was sought from the patient-led Kings Electronic Records Research Committee (KERRI) which operates under an opt-out model under London South East Research Ethics Committee approval (reference 18/LO/2048).

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

DATA AVAILABILITY

Only aggregated counts over time is available (no patient-level data due to privacy). Any request for this will need to be reviewed first by the hospital information governance committee which includes the Caldicott Guardian and Data Protection Officer.

CODE AVAILABILITY

Source code for the platform is available at (https://github.com/CogStack).

Received: 2 September 2020; Accepted: 26 January 2021;
Published online: 24 February 2021

REFERENCES

1. Ginsberg, J. et al. Detecting influenza epidemics using search engine query data. *Nature* **457**, 1012–1014 (2009).
2. Lazer, D., Kennedy, R., King, G. & Vespignani, A. The parable of Google Flu: traps in big data analysis. *Science* **343**, 1203–1205 (2014).
3. Hodson, H. [dn25217] Google Flu Trends gets it wrong three years running. *N. Sci.* **221**, 24 (2014).
4. Lacey, K. Google Flu. *Encycl. Big Data* https://doi.org/10.1007/978-3-319-32001-4_106-1 (2017).
5. Estiri, H. et al. Transitive sequencing medical records for mining predictive and interpretable temporal representations. *Patterns* **1**, 100051 (2020).
6. Araz, O. M., Bentley, D. & Muelleman, R. L. Using Google Flu Trends data in forecasting influenza-like-illness related ED visits in Omaha, Nebraska. *Am. J. Emerg. Med.* **32**, 1016–1023 (2014).
7. Choi, S. B. & Ahn, I. Forecasting seasonal influenza-like illness in South Korea after 2 and 30 weeks using Google Trends and influenza data from Argentina. *PLOS ONE* **15**, e0233855 (2020).
8. Wada, K. Correlation of 'Google Flu Trends' with sentinel surveillance data for influenza in 2009 in Japan. *Open Public Health J.* **4**, 17–20 (2011).
9. Yang, S. et al. Using electronic health records and Internet search information for accurate influenza forecasting. *BMC Infect. Dis.* **17**, 1–9 (2017).
10. Timpka, T. et al. Performance of eHealth data sources in local influenza surveillance: a 5-year open cohort study. *J. Med. Internet Res.* **16**, e116 (2014).
11. Menni, C. et al. Real-time tracking of self-reported symptoms to predict potential COVID-19. *Nat. Med.* **26**, 1037–1040 (2020).
12. Jackson, R. et al. CogStack – experiences of deploying integrated information retrieval and extraction services in a large National Health Service Foundation Trusthospital. *BMC Med. Inform. Decis. Mak.* **18**, 47 (2018).

ACKNOWLEDGEMENTS

The infrastructure and the authors received funding from: National Institute for Health Research (NIHR) Biomedical Research Centre at South London and Maudsley NHS Foundation Trust, Health Data Research UK, UK Research and Innovation, London Medical Imaging & Artificial Intelligence Centre for Value Based Healthcare, Innovate UK, the NIHR Applied Research Collaboration South London, the NIHR University College London Hospitals Biomedical Research Centre and King's College London. We also thank all patients who agreed to their data being used for research, and the patients of the KERRI committee.

AUTHOR CONTRIBUTIONS

J.T., R.D., W.B. conceived the study design. V.D., P.D., V.O., J.T., R.D. performed data processing and architecture development. J.T., V.D., R.D. wrote the manuscript. V.D., P.D., V.O., J.T., R.B., C.B., R.D. performed critical review.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41746-021-00406-7.

Correspondence and requests for materials should be addressed to J.T.H.T.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.