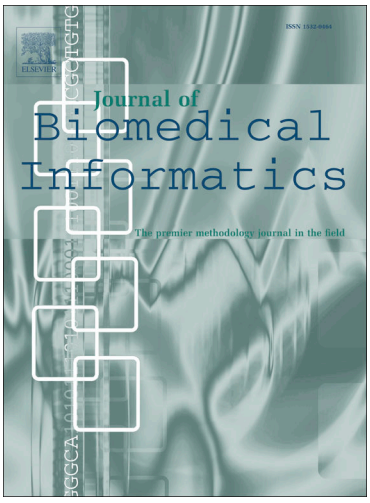
Article:
Goodday, S.M., Kormilitzin, A., Vaci, N. et al. (2020) Maximizing the use of social and behavioural information from secondary care mental health electronic health records. Journal of Biomedical Informatics, 107. 103429. ISSN: 1532-0464

https://doi.org/10.1016/j.jbi.2020.103429

Journal Pre-proofs



Maximizing the use of social and behavioural information from secondary care mental health electronic health records

S.M. Goodday, A. Kormilitzin, N. Vaci, Q. Liu, A. Cipriani, T. Smith, A. Nevado-Holgado

PII: S1532-0464(20)30057-5
DOI: https://doi.org/10.1016/j.jbi.2020.103429
Reference: YJBIN 103429

To appear in: *Journal of Biomedical Informatics*

Received Date: 6 January 2020
Revised Date: 15 April 2020
Accepted Date: 19 April 2020

Please cite this article as: Goodday, S.M., Kormilitzin, A., Vaci, N., Liu, Q., Cipriani, A., Smith, T., Nevado-Holgado, A., Maximizing the use of social and behavioural information from secondary care mental health electronic health records, *Journal of Biomedical Informatics* (2020), doi: https://doi.org/10.1016/j.jbi.2020.103429

Maximizing the use of social and behavioural information from secondary care mental health electronic health records

Goodday SM[1,2], Kormilitzin A[1], Vaci N[1], Liu Q[1], Cipriani A[1], Smith T[3], Nevado-Holgado A[1].

[1]Department of Psychiatry, University of Oxford
[2]4youandme, Seattle, Washington, USA
[3] Oxford Health NHS Foundation Trust

Corresponding author:

Sarah M. Goodday, PhD
Department of Psychiatry
University of Oxford
Warneford Hospital
Warneford Ln, Oxford OX3 7JX
Sarah.goodday@psych.ox.ac.uk

4youandme.org

Maximizing the use of social and behavioural information from secondary care mental health electronic health records

Word count: 2678

Abstract: Purpose: The contribution of social and behavioural factors in the development of mental health conditions and treatment effectiveness is widely supported, yet there are weak population level data sources on social and behavioural determinants of mental health. Enriching these data gaps will be crucial to accelerating

(EHR) as a source of non-clinical determinants, although social and behavioural information are not systematically collected metrics in EHRs, internationally. Objective: In this commentary, we highlight the nature and quality of key available structured and unstructured social and behavioural data using a case example of value counts from secondary mental health data available in the UK from the UK Clinical Record Interactive Search (CRIS) database; highlight the methodological challenges in the use of such data; and possible solutions and opportunities involving the use of natural language processing (NLP) of unstructured EHR text. Conclusions: Most structured non-clinical data fields within secondary care mental health EHR data have too much missing data for adequate use. The utility of other non-clinical fields reported semi-consistently (e.g., ethnicity and marital status) is entirely dependent on treating them appropriately in analyses, quantifying the many reasons behind missingness in consideration of selection biases. Advancements in NLP offer new opportunities in the exploitation of unstructured text from secondary care EHR data particularly given that clinical notes and attachments are available in large volumes of patients and are more routinely completed by clinicians. Tackling ways to re-use, harmonize, and improve our existing and future secondary care mental health data, leveraging advanced analytics such as NLP is worth the effort in an attempt to fill the data gap on social and behavioural contributors to mental health conditions and will be necessary to fulfill all of the domains needed to inform personalized interventions.

Keywords: electronic health records; natural language processing; precision medicine; mental health; selection bias; data quality

Introduction

The importance of social and behavioural determinants on mental health outcomes is widely supported(1-3), yet, we lack robust data sources in these domains. Mental health outcomes vary considerably by ethnicity, culture, gender, education, income, family and early adversity and influence not only mental health disease burden, but healthcare delivery and service pathways(1, 3). Behavioural factors such as smoking, alcohol use and exercise are among the most robust modifiers of mental health conditions(2), and are critically important for effective interventions as these factors exacerbate disease and contribute to morbidity. The contribution of these contextual domains is recognized in the definition of precision medicine, which can also apply to mental health - to incorporate genetic, biological, behavioural and environmental data from individuals to enable tailored interventions and clinical decision making(4). Originally highlighted by the WHO landmark report 'Closing the Gap'(3), we still have weak population level data on social determinants of health; therefore, enriching this data gap should be a priority. Electronic Health Records (EHRs) electronically store health information comprising longitudinal patient trajectories across multiple visits within the health care system. Social and behavioural metrics are not routinely collected within primary and secondary care health settings in the UK or internationally with exception to a small number of behavioural and socio-demographic factors that are screened. The potential integration of social or behavioural information into EHR systems will be important if we intend to use EHR data for precision medical research(5-7). Although, barriers to this implementation include that clinic appointment times are short, EHR reporting can cause disruptions in workflow and clinician burnout is prevalent(8). The use of EHR data was not originally intended for secondary analysis, and particularly not for non-clinical data, yielding challenges in its use relating to missing data and associated biases.

However, EHR data are increasingly being used for such purposes. Further, most fields in EHR systems are unstructured(9). Fortunately, natural language processing (NLP), a subfield of linguistics and computer science focused on the interactions between

opportunity to extract rich contextual information from clinical notes and attachments, particularly from secondary care mental health service data.

With recent efforts in the larger scale, federated use of EHR data(10), understanding the availability, utility and caveats of the available service data is of importance for those considering use of these data. EHR systems are complex, and their use must be done with caution to fully understand the intricacy of the system from how and why clinical data are reported in the front end, to how this information translates into structured and unstructured fields in the back end. In this paper, we highlight the nature and quality of key available structured and unstructured non-clinical fields through value counts in secondary care mental health data available in the UK and highlight the complexity and implications of their use and the promise of NLP to maximize their use.

UK-CRIS data

As an example of available non-clinical data from EHRs, we report value counts of key data fields from UK Clinical Record Interactive Search (UK-CRIS) at Oxford Health NHS Foundation Trust (CRIS-Oxford), one of the 12 NHS Mental Health Trusts participating in the CRIS programme. CRIS-Oxford reflects data reported in secondary mental health service settings from the Oxfordshire, Buckinghamshire, Swindon and Wiltshire region (Table 1). UK-CRIS is the primary programme in the UK dedicated to creating anonymized secondary care mental health data comprising >2 million patient EHRs (https://crisnetwork.co/). This programme provides a mechanism for researchers to access secondary care mental health data from all patients EHRs within each participating NHS Mental Health Trust. The accessible data is translated directly from the patient EHR into structured text (e.g., diagnosis, event, referral, discharge and administrative codes) and 80% of the available unstructured text (medical notes and attachments). While there are federated approaches to accessing data across the 12 participating NHS Mental Health Trusts, accessing data from all groups simultaneously is still a challenge, and therefore the representativeness of patient records from CRIS-Oxford compared to other NHS Trusts is unknown.

Social and behavioural structured text data

Not surprisingly, data reporting on gender, ethnicity, refugee status and truncated postcode is acceptable. Although less systematically reported, fields including religion, country of origin, first language and marital status have considerable numbers of patients with data, keeping in mind these fields would typically be reported in adults. Several potentially useful variables have less than 1000 patients with data indicating less systematic reporting. However, the reporting process for several of these variables

is restricted by certain contexts. For example, there are only carer assessment data on patients with carers, while information on school is only relevant in children and adolescents.

The range of lifestyle information currently available from CRIS-Oxford is limited. Information on smoking and alcohol can be retrieved by searching different tables such as admission checklists for adults, from the Car Relax Alone Forgot Friends Trouble (CRAFFT) screening test scores for substance use problems in adolescents, or from mental health and emergency departments assessments and are present in both structured and unstructured fields. Ironically, a recent Commissioning for Quality and Innovation (CQUIN)(11) was commissioned in March, 2019 to increase the systematic screening for smoking and alcohol use in adult inpatient settings including mental health services indicating that the quality of these variables may improve in the near future.
A challenge in interpreting these data is in the lack of clarity of what constitutes routinely collected clinical information compared to information not routinely collected that might vary both within and across NHS Trusts. Certain pieces of information are collected at different times along patient journeys (some only within emergency department visits) and can be found in different locations within EHRs with different numbers of patients. Further, we know that clinicians are differentially reporting this information, but we don't necessarily know why. However, evidence suggests that demographic factors and disease status influence the likelihood of clinician reporting of specific health outcome data(12, 13).

Unstructured text data from clinical notes and attachments

Clinical notes are clinician written notes surrounding ward round notes, phone calls, clinical observations, while attachments can reflect letters from GPs, test results, referral and clinic letters or reports. These unstructured data contain rich information about patients at the symptom level as well as context, that is not captured through diagnostic classification codes or other rigid structured fields. In the context of non-clinical data, structured fields do not often allow sufficient detail to describe the social or environmental contexts relating to patients experience of disease. Particularly given that clinicians are still using text-based methods to document health-related information as the volume of unstructured text data largely surpasses structured data(9).

The promise of Natural Language Processing (NLP) of unstructured EHR text

One of the main objectives of an information extraction system is the identification of concepts of interests in free-text (i.e. drug names, symptoms or events), known as named-entity recognition (NER) and further classification of relationships between the identified entities(14, 15, 16). Traditionally, such systems were implemented using logical rules and various pattern matching algorithms, consulting large dictionaries. More recently, the field of NLP has seen an explosion of advanced methods, such as vector representation of words, an application of deep neural networks to NLP-related

tasks, transfer learning and Transformer architecture, which paved the way towards novel and large Transformer-like language models to learn contextual information from texts(15), and in particular from EHR unstructured texts(17). An accurate annotated

done by field experts (i.e., clinicians, nurses), who are able to recognise concepts (i.e., symptoms, adverse events, behaviours) and mark corresponding text spans with appropriate labels. The NER model learns patterns from the annotated examples and then will be able to identify such patterns in new, unseen data(17, 18).

A more recent standard in the NLP community due to its higher accuracy and robustness when compared to earlier machine learning algorithms is the deep Bidirectional Transformers for Language Understanding **(**BERT) model(19). These models transform free text into ordered sequences of numerical vectors that are fed to a series of neural network layers with embedded weights to define their behavior. Once the information extraction pipeline is trained and optimised, the unstructured clinical notes are parsed and the concepts of interest are extracted, labelled and organised in a tabulated format (i.e. XML or CSV file). The identified concepts are organised in tuples with relevant attributes. For example, the sentence: *"The patient was switched from drug A to drug B 10mg nocte for 30 days"* will be parsed into two tuples with the information related to medications: {"Drug": "drug A", "Assertion": "stopped", "Strength": Null, "Frequency": Null, "Duration": Null} and {"Drug": "drug B", "Assertion": "affirmed", "Strength": "10mg", "Frequency": "nocte", "Duration": "30 days"}. Such organised tuples, augmented with temporal information available from the texts, represent the chronological progression of events in the patients' history and could be used in statistical and machine learning analyses. While this example pertains to treatment, one could imagine the opportunity in other areas of information extraction.

Available clinical notes and attachments from CRIS-Oxford

Attachment and clinical note category types from CRIS-Oxford data provides close to 70 unique types. These types were selected from an available value list that best summarises the note or attachment content. There is a wide range of different unstructured text documents covering different aspects of the patient experience ranging from assessments and reports on clinical status to reports on social care circumstances to risk behaviours and adverse events such as abuse or assault (Table 2). Much of the focus on the interrogation of EHR data using NLP has focused on clinical symptoms, but the range of information from the available attachments and clinical notes spans much farther than this domain and is deserving of future research into the feasibility of extracting socio-demographic, lifestyle and environmental information.

Limitations

Most structured non-clinical data fields within secondary care mental health EHR data available from CRIS-Oxford have too much missing data for adequate use. There are a few useful structured non-clinical variables beyond those reported consistently in a

their use depends entirely on treating them appropriately in analyses and carefully interpreting the many reasons behind missing data. There are several challenges associated with the use of non-clinical information from EHR data(20). When using structured variables that are not systematically reported, understanding the many different types of missing data that are often missing-not-at-random imposing selection bias is crucial. Missingness could reflect true absence (a real zero), patient refused to provide information, clinician did not ask patient, or does not report on the particular variable for 'x' reason, missing due to de-identification for confidentiality purposes, or not relevant due to age or other contexts. Quantifying the volume of important "missingness" in the context of the appropriate denominator to quantify selection bias is recommended. Imputation is often employed to deal with missingness, yet should be used with caution as these methods can often lead to distorted estimates, while most methods are unable to adequately account for selection bias(21).

New opportunities exist in the exploitation of unstructured text from secondary care EHR data. While there are challenges in the use of NLP relating to subjectivity biases in the expressiveness and perceptions of clinicians in how they describe and report information, these current limitations are all active areas of research in the NLP field(22). An additional limitation that can affect the development of a robust and accurate information extraction system is the need for a large amount of high quality manually annotated data. The availability of a sufficient amount of annotated medical concepts also affects the level of granularity, at which a NER system may recognise and disambiguate the concepts. For example, the recognition of well-defined categories, such as drug names or routes of administration of medications requires less annotated data as compared to differentiation between the levels of educations or a history of parental substance abuse, which may be presented in a highly variable way. Possible approaches to address the problem of training an information extraction model from insufficient amount of annotated data include transfer learning paradigms, where language models are first pre-trained on large collections of publicly available texts such as Wikipedia or Common Crawl corpora. It is also worth mentioning, that some medical concepts are intrinsically underreported by clinical practitioners in medical records, and NLP approaches cannot overcome this inconsistent reporting.

Potential solutions

A top down approach reconsidering how we report health-related indicators from signs and symptoms to determinants and contributors to mental health outcomes could improve data quality. Existing time constraints by clinicians can ultimately be overcome by a cultural change in how health professionals and patients see routinely collected

data and by restructuring how medical information is recorded(5, 7). Increasing transparency on what is routinely reported versus not would lend useful information for the interpretation of missing data from EHR systems. Universal screening for social

recent years(13) (while out of scope of this commentary) is deserving of research into the acceptability from both clinicians and patients. Universal screening would importantly accommodate the primary limitation of inconsistent reporting in using EHR data.

The inclusion of more systematic patient reported outcome measures, or participant generated health data could further close this data gap. The advancements and ubiquity of digital technology offers opportunities for alternative data capture outside clinic visits. A window of opportunity here is the True Colours Remote Symptom Monitoring System, a web-based and SMS-based application for daily and weekly symptom monitoring that is currently used by a considerable number of patients with psychiatric conditions in the Oxfordshire and Buckinghamshire regions(23). Digital systems such as True Colours could serve as a platform for high-frequency active and passive lifestyle measurements(24,25).

Linkage studies of other data sources that contain social and behavioural data to EHRs is an alternative solution. However, most population data sources reflect cross-sectional surveys, or lack the necessary identifiers to link these sources to EHRs. While the UK-Biobank has been linked to UK-CRIS data, this survey lacks robust data on social and lifestyle factors and is limited to older adults. A unique example involves the prospect of linking education and social care data to UK-CRIS, which has been done(26), offering exciting avenues to explore early childhood risk factors of mental health trajectories. To date, these linkage instances in the UK are rare owing to complex data governance and political challenges.

The rich information from unstructured secondary care mental health data reflects a timeline of a patient's symptoms, medications, adverse events, but also potentially a chronological timeline into what is currently happening in their lives that has resulted in a visit to the doctor. The uniqueness of UK-CRIS is in providing access to the unstructured EHR text to researchers, providing ample opportunity to exploit a wide range of information, not presently captured or availability at the structured level. Novel signature-based machine learning methods are being developed to make use of such rich time-series data that transform the sequential data into useful features which feed into algorithms to identify robust combinations predictive of outcomes(27-29). The downstream potential for embedded models within EHR systems is to translate unstructured text into a structured format comprising clinically meaningful variables, and important determinants of mental health. Making this information accessible by standard statistical software packages and for use in downstream analyses that would be a huge benefit to the wider scientific community.

Conclusions

Tackling ways to re-use, harmonize, and improve our existing and future secondary

in an attempt to fill the data gap on social and behavioural contributors to mental health conditions. Understanding the characteristics of these data, their limitations in use and their potential for health research is of paramount importance. This effort should be collaborative and international reflecting opportunities to improve outcomes in patients by shifting focus to factors that contribute to or exacerbate disease pathways (prevention), rather than the sole focus on acute care and treatment.

Funding: This project was supported by the UK Clinical Record Interactive Search (UK-CRIS) system funded by the National Institute for Health Research (NIHR) and the Medical Research Council, with the University of Oxford, using data and systems of the NIHR Oxford Health Biomedical Research Centre (BRC-1215-20005).

AC is supported by the National Institute for Health Research (NIHR) Oxford Cognitive Health Clinical Research Facility, by an NIHR Research Professorship (grant RP-2017-08-ST2-006) and by the NIHR Oxford Health Biomedical Research Centre (grant BRC-1215-20005).

SG, AK, NV and QL are supported by the MRC under the Pathfinder programme grant MC/PC/17215.

The views expressed are those of the authors and not necessarily those of the UK National Health Service, the NIHR, or the UK Department of Health.

Acknowledgements:

We would also like to acknowledge the work and support of the Oxford CRIS Team with conducting the value counts in relation to this work, Adam Pill and Suzanne Fisher, CRIS Academic Support and Information Analysts.

References

1. Allen J, Balfour R, Bell R, Marmot M. Social determinants of mental health. Int Rev Psychiatry. 2014;26(4):392-407.
2. Walsh R. Lifestyle and mental health. Am Psychol. 2011;66(7):579-92.
3. Marmot M, Friel S, Bell R, Houweling TA, Taylor S, Commission on Social Determinants of H. Closing the gap in a generation: health equity through action on the social determinants of health. Lancet. 2008;372(9650):1661-9.
4. Tomlinson A, Furukawa TA, Efthimiou O, Salanti G, De Crescenzo F, Singh I, et al. Personalise antidepressant treatment for unipolar depression combining individual choices, risks and big data (PETRUSHKA): rationale and protocol. Evid Based Ment Health. 2019.
5. Gottlieb LM, Tirozzi KJ, Manchanda R, Burns AR, Sandel MT. Moving electronic medical records upstream: incorporating social determinants of health. Am J Prev Med. 2015;48(2):215-8.
6. Cantor MN, Thorpe L. Integrating Data On Social Determinants Of Health Into Electronic Health Records. Health Aff (Millwood). 2018;37(4):585-90.
7. Andermann A. Screening for social determinants of health in clinical care: moving from the margins to the mainstream. Public Health Rev. 2018;39:19.
8. Medisauskaite A, C. K. Does occupational distress raise the risk of alcohol use, binge-eating, ill health and sleep problems among medical doctors? A UK cross-sectional study. BMJopen. 2019;9:e027362.
9. Murdoch TB, Detsky AS. The inevitable application of big data to health care. JAMA. 2013;309(13):1351-2.
10. Riley WT, Nilsen WJ, Manolio TA, Masys DR, Lauer M. News from the NIH: potential contributions of the behavioral and social sciences to the precision medicine initiative. Transl Behav Med. 2015;5(3):243-6.
11. NHS. Guidance Health Matters: tobacco and alcohol CQUIN. 2019.
12. Petersen I, Welch CA, Nazareth I, Walters K, Marston L, Morris RW, et al. Health indicator recording in UK primary care electronic health records: key implications for handling missing data. Clin Epidemiol. 2019;11:157-67.
13. Davidson KW, McGinn T. Screening for Social Determinants of Health: The Known and Unknown. JAMA. 2019.
14. Hofer M KA, Goldberg P, Nevado-Holgado A. Few-shot Learning for Named Entity Recognition in Medical Text. arXiv. 2018;preprint arXiv:1811.05468.
15. Gligic L KA, Goldberg P, Nevado-Holgado A. Named entity recognition in electronic health records using transfer learning bootstrapped neural networks. Neural Networks. 2020;1(121):132-9.

16. Kormilitzin, A., Vaci, N., Liu, Q. and Nevado-Holgado, A., 2020. Med7: a transferable clinical natural language processing model for electronic health records. *arXiv preprint arXiv:2003.01271*.
17. Senior, M., Burghart, M., Yu, R., Kormilitzin, A., Liu, Q., Vaci, N., Nevado-Holgado, A., Pandit, S., Zlodre, J. and Fazel, S., 2020. Identifying predictors of suicide in severe mental illness: a feasibility study of a clinical prediction rule (OxMIS). *European Psychiatry*, *63*(1).

18. Vaci N, Liu Q, Kormilitzin A, De Crescenzo F, Kurtulmus A, Harvey J, et al. Natural language processing for structuring clinical text data on depression using UK-CRIS. Evid Based Ment Health. 2020;23(1):21-6.
19. Devlin, J., Chang, M.W., Lee, K. and Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
20. Hollister B aBV. Should Electronic Health Record-Derived Social and Behavioral Data Be Used in Precision Medicine Research? AMA Journal of Ethics. 2018;20(9):E873-80.
21. Beaulieu-Jones BK, Lavage DR, Snyder JW, Moore JH, Pendergrass SA, Bauer CR. Characterizing and Managing Missing Structured Data in Electronic Health Records: Data Analysis. JMIR Med Inform. 2018;6(1):e11.
22. Meystre SM, Savova GK, Kipper-Schuler KC, Hurdle JF. Extracting information from textual documents in the electronic health record: a review of recent research. Yearb Med Inform. 2008:128-44.
23. Goodday SM AL, Goodwin G, Saunders K, South M, Mackay C, Denis M, Hinds C, Attenburrow J, Davies J, Welch J, Stevens W, Mansfield K, Suvilehto K, Geddes J. The True Colours remote symptom monitoring system: A decade of evolution. Journal of Internet Medical Research. 2019;In Press.
24. Walsh, A., Kormilitzin, A., Hinds, C., Sexton, V., Brain, O., Keshav, S., Uhlig, H., Geddes, J., Goodwin, G., Peters, M. and Collins, G., 2019. Defining faecal calprotectin thresholds as a surrogate for endoscopic and histological disease activity in ulcerative colitis—a prospective analysis. *Journal of Crohn's and Colitis*, *13*(4), pp.424-430.
25. Kormilitzin, A., Walsh, A.J., Matini, L., Kantschuster, R., Lepetyukh, M., Wilson, J., Brain, O., Palmer, R., Ambrose, T., Satsangi, J. and Travis, S.P.L., Patient-reported symptoms over a period of 14 days reliably predict endoscopic and histological disease activity in ulcerative colitis (UC). *Journal of Crohn's and Colitis*. 2020; *14*
26. Downs JM, Ford T, Stewart R, Epstein S, Shetty H, Little R, et al. An approach to linking education, social care and electronic health records for children and young people in South London: a linkage study of child and adolescent mental health service data. BMJ Open. 2019;9(1):e024355.
27. Morrill J KA, Nevado-Holgado A, Swaminathan S, Howison S, Lyons T. The Signature-based Model for Early Detection of Sepsis from Electronic Health Records in the Intensive Care Unit. Computing in Cardiology Conference (CinC). 2019;IEEE.
28. Kormilitzin AB, Saunders, K. E. A., Harrison, P. J., Geddes, J. R., & Lyons, T. J. Application of the signature method to pattern recognition in the CEQUEL clinical trial. arXiv preprint arXiv:160602074. 2016.
29. Kormilitzin A, Saunders, K. E., Harrison, P. J., Geddes, J. R., & Lyons, T. Detecting early signs of depressive and manic episodes in patients with bipolar disorder using thesignature-based model. arXiv preprint arXiv:170801206. 2017.

Table 1. Key social and behavioural data available in CRIS-Oxford secondary care mental health data (total number of patients: 158,000 patients; 38,000<18 years of age)

CRIS category	Available field	Patients with data[h]	Nature of variable
Patient Information – socio-demographic	Gender	>95%	Structured
	First language[a]	15%	Structured
	Religion[a]	20%	Structured
	Country of origin[a]	10%	Structured
	Ethnicity[a]	80%	Structured
	Refugee status	>95%	Structured
	Truncated postcode	>95%	Structured
	School	<5%[i]	Unstructured
	Sexuality[b]	<5%[j]	Structured
Clinic assessments – cultural needs	Faith importance	<5%	Structured
	Chaplain value	<5%	Structured
	Faith belief	<5%	Unstructured
Patient information – family circumstance	Martial status[c]	45%[j]	Structured
	Lives with[d]	<5%	Structured
	Family legal status[e]	<5%	Structured
	Family parental responsibility	<5%[j]	Structured
Carers assessment	Person safe, presence of emotional support, dealing with crisis, presence of abuse/violence, worry, life management	<5%	Structured
	Smoking[f]	20%	Structured/ unstructured
	Alcohol[g]	10%	Structured/ unstructured
	CAMHS_CRAFFT	10%[i]	Structured/ unstructured
	Nutrition assessment	<5%	Structured/ unstructured

[a]Includes several categories (85+) requiring re-coding

[b]Gay, lesbian, bisexual, gay/lesbian, heterosexual, not stated (person asked but declined to ask a response), person asked b99ut does not know or is not sure)
[c]Civil partnership, co-habitating, divorced, married, married/civil partner, separated, single, widowed, not applicable, not known, not disclosed

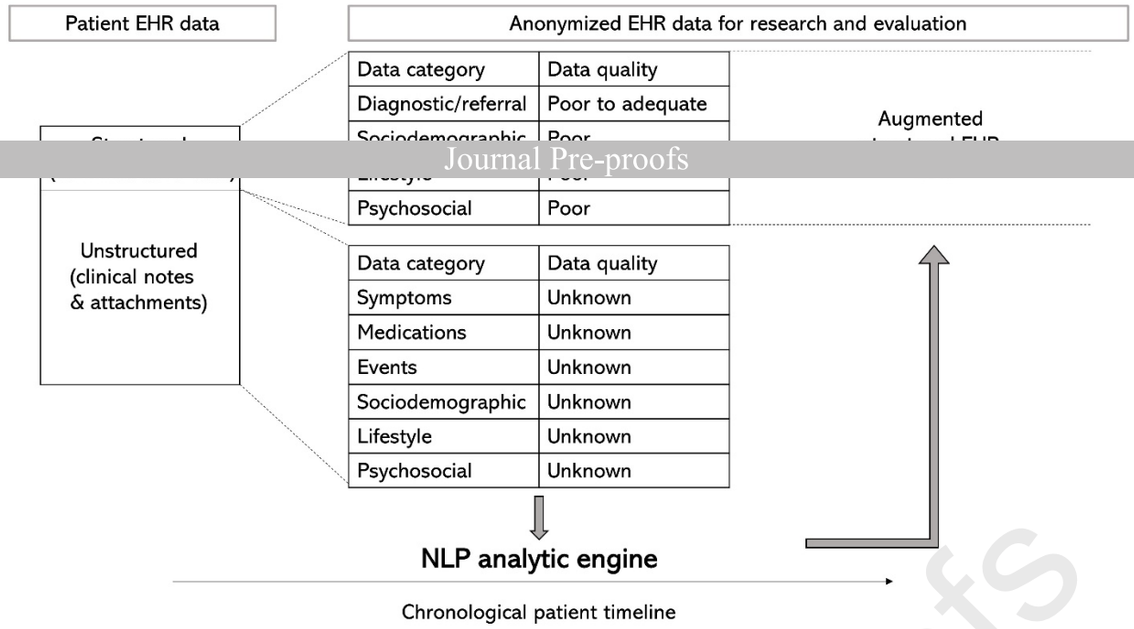
known, other
[e]Informal care, section 20 – accommodated, section 38 – interim care order, unknown
[f]From CRIS tables (CPA review, CPA start, Adult admission checklist, mental health assessment, physical health review, ED assessment)
[g]From CRIS tables (CAMHS_CRAFFT, ED assessment, mental health assessment)
[h]Please interpret with caution. These numbers reflect value counts that are reported under differential contexts. These values cannot be equated to EHR health indicator reporting within these regions
[i]Denominator only includes patients <18 years
[j]Denominator only includes patients ≥18 years
Variables presented were last populated within CRIS-Oxford in April 2019. Date when variables were first populated ranges from April, 2015 to June, 2017

Table 2. Attachments and clinical note category types from CRIS Oxford secondary care mental health data

Attachments	**Clinical notes**
Advance directives/Living will	7 day follow up
Reports/Assessment	Abuse
Report from Social Services	Accident
Social Care Commissioning	Aggressive
Care Plans and requests	Assault Actual - Perpetrator
Clinical Notes and summaries	Assault Actual - Victim
Deprivation of Liberty Safeguards	Assault Threat - Perpetrator
Discharge summaries	Assault Threat - Victim
Letters/referrals	CPA - Discharge
Information from Patient	CPA - Initial
Inpatient communication feedback	CPA - Ongoing
Investigation and Test results	Default notes
Lasting Power of Attorney	Deliberate Self Harm
Medication	Follow Up
Mental Capacity Act	General Update
Documents relating to the MHA	Handover
Section 17	Home Visit
Observations	Initial Assessment
PCIS	Looked After Children
Printable CPA Review	Meaningful Activity
Safeguarding	Medication
Sensitive and 3rd Party Information	MHA assessment
	Neglect of others
	Nurse
	OT
	Other Risk Behaviours
	Phlebotomy
	Psychology
	Review
	Safeguarding
	Suicide

	Self Harm
	Self Harm Actual
	Self Harm Threat
	Self Neglect
	Social Worker
	Therapy
	Urgent Assessment
	Violent
	Vulnerability From Others
	Ward Review

CPA: care program approach; DNACPR: Do Not Attempt Cardio-pulmonary Resuscitation; MHA: mental health act; OT: occupational therapy; PCIS: primary care information service; Section 17: Provision that a responsible clinician may grant a detained patient leave of absence from a hospital



Highlights

- Precision medicine will require better data sources of social and behavioural data
- Most structured social and behavioural data fields from EHR data are inadequate
- NLP of unstructured EHR text could yield psychosocial timeline of risk data
- Social and behavioural information could be extracted from unstructured EHR data