phuse·virtual

Single Day Event

Ameliorate Data Quality in a Clinical Registry by Natural Language Processing

Sandhya Pitta
PAREXEL International

Disclaimer

All opinions expressed in this presentation are the presenter's personal views, and do not reflect the views or opinions of Parexel International

Content

- Introduction
- Frame work of the automated verification approach

  I) Paper-based CRFs' Recognition

  II) EMR data extraction

  III) Automated verification implementation

- Case Study: Chinese Coronary Artery Disease
- References

Introduction

- A clinical registry is defined as a collection of information about individuals, usually for a public health and focused around a specific diagnosis or condition.

- The quality of data collection is crucial in registry studies, it has an impact on registry study credibility and is emphasize by clinical researchers.

- Good Clinical Practice requires that data must be accurate, complete and verifiable to ensure patient safety and data quality.

- Electronic case report forms (**eCRFs**), which guarantee the quality of data collection to a certain degree, are employed in some registry studies. Paper-based CRFs are still used in many registry studies as they better fit busy clinical environment and traditional working habit of clinicians.

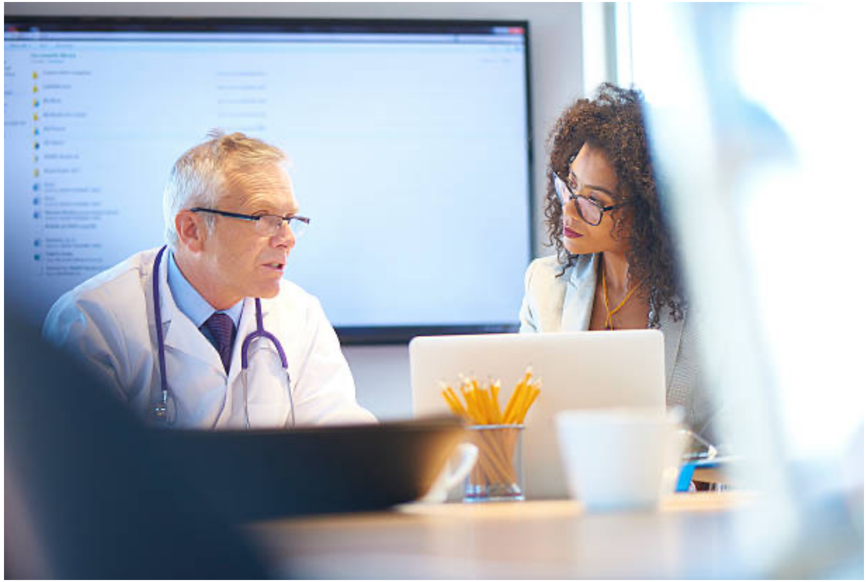
Tian et al., 2019

Introduction

- ➢ The transcription procedure of data from paper-based CRFs to eCRFs is error-prone

- ✓ **Double data entry:** **Initial entry step** and **a verification step**; each step is performed by a separate data entry clinician. However, it is time-consuming, labor-intensive and ineffective for detecting recording errors.

- ✓ **Source data verification (SDV):** A verification of conformity of the data recorded in case report forms with source.



Tian et al., 2019

Introduction

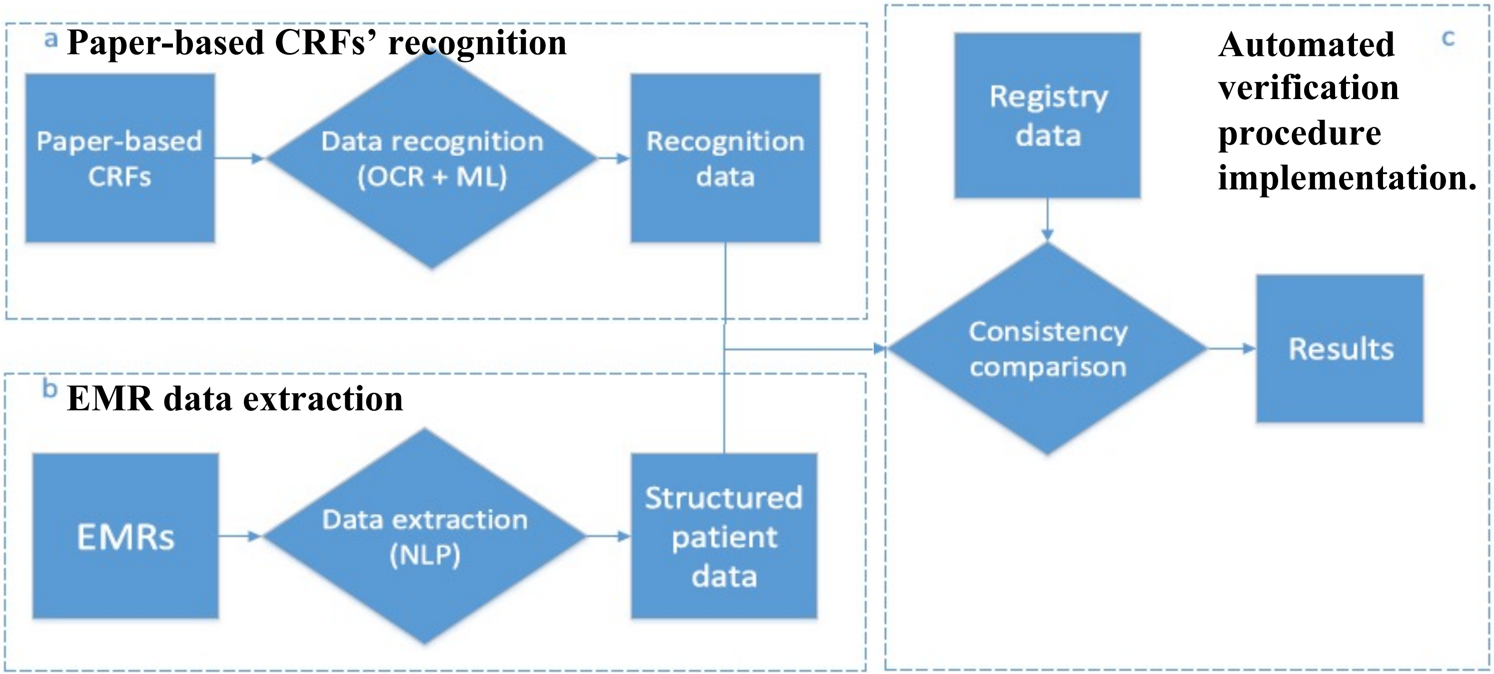
- Due to the restrictions in applying an electronical data capture system and traditional work habit of data recording are still using paper- based CRFs and then transcribe them into an eCRFs.

- For data verification, base original data is not enough to improve data quality of registry data because it only reveals data errors happening in the data transfer procedure. Using multiple data sources like original paper-based CRFs and EMRs could enhance the effect of verification and discovery for more error or missing data.

- To verify registry data they used automated verification approach such as Machine learning enhanced optical character recognition (OCR) and Natural Language Processing (NLP) techniques.

Tian et al., 2019

Framework of the automated verification approach



Tian et al., 2019

Paper-based CRFs' recognition

DOCUMENT SCAN

SCANNED IMAGE FILE

Scan the paper-based CRFs to transform to image files in PNG format

OCR (Optical Character Recognition)

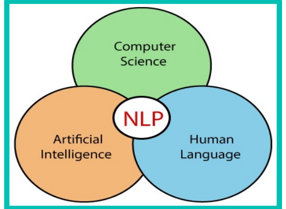
Segment scanned images into smaller pictures where the content need to be recognized with OCR.

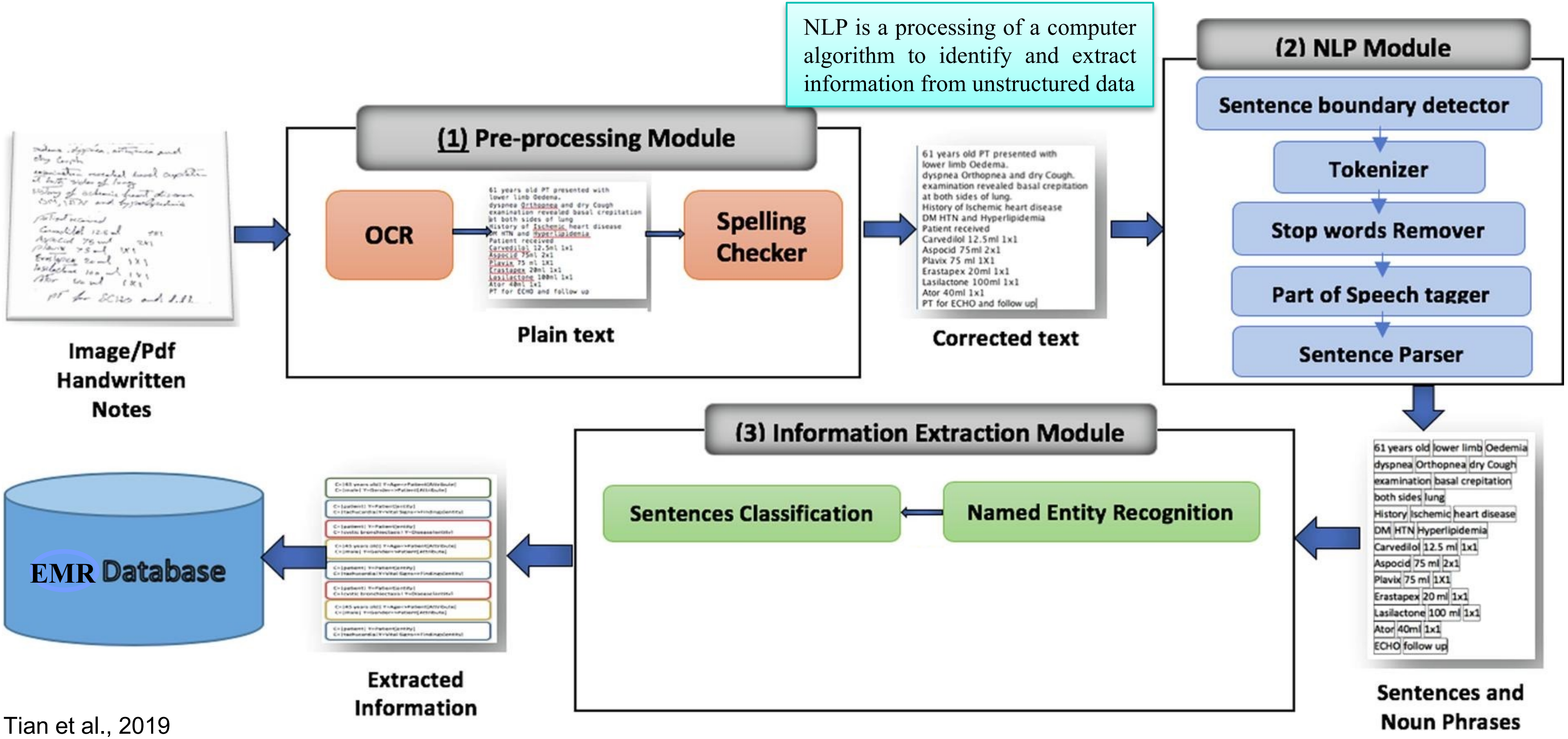
TEXT DOCUMENT

- Marks of checkboxes (with Machine learning)
- Hand-written numbers (with convolution neural network model)

Tian et al., 2019

EMR data extraction





Tian et al., 2019

Automated verification procedure implementation

Select the value of the data field from different data sources.

Compare whether these values are consistent with each other.

JAVA language to verify data automatically by retrieving values of one data field from different data source based on application interfaces or query language. Then, this procedure compares the consistency of those data values and synthesizes the results accordingly

Give a conclusion based on the result of the previous step.

Data errors possibly occurring in a data collection are of four types:

Entry data missing: data missing in transcribed data from paper-based CRF to registry database

Entry data Errors: data error in transcribed data from paper-based CRF to registry database

Recording data Missing: data missing in recording data on paper-based CRF

Recording data Errors: data error in recording data on paper-based CRF

Tian et al., 2019

Possible conclusion of different data existence status and consistency situations

	CRF data existence status	EMR data existence status	Consistency situation	Conclusion
Registry data is null	False	False	Data not exist	/
	False	True	/	Recording data missing
	True	False	/	Entry data missing
	True	True	CRF data consistent with EMR data	Entry data missing
			CRF data inconsistent with EMR data	Entry data missing
Registry data not null	False	False	/	Entry data error
	False	True	Registry data consistent with EMR data	Entry data error
			Registry data inconsistent with EMR data	Entry data error
	True	False	Registry data consistent with CRF data	Correct
			Registry data inconsistent with CRF data	Entry data error
	True	True	Data element consistent with each other	Correct
			Registry data consistent with CRF data, but inconsistent with EMR data	Recording error or correct, more data analysis is needed.
			CRF data consistent with EMR data, but inconsistent with registry data	Entry error
			Registry data inconsistent with CRF data, but consistent with EMR data	Entry error
			Data element inconsistent with each other	Recording error or correct, more data analysis is needed

Tian et al., 2019

Case Study: Chinese Coronary Artery Disease

- Automated approach has applied to a **Chinese Coronary Artery Disease** Registry (CCADR) to assess the feasibility of the approach
- In this study they identifying data quality errors, designed an experiment to evaluate the effectiveness and accuracy of the automated approach and manual verification
- For this Study they selected **150** enrolled patients randomly from the registry. Then, invited two professional data experts to verify the data of the **150** patients carefully and one clinical expert confirmed the results. the verification results of experts as the gold standard in experiment
- Selected three clinicians who are relevant to the registry topic to verify the data of the same 150 patients manually. Then, used the automated approach to verify the same data
- Finally, calculated the accuracy and recall of the average manual approach and the automated approach

Tian et al., 2019

Case Study: Chinese Coronary Artery Disease

Data corresponding results between registry data and EMRs.

Data fields category	Amount of data fields	Corresponding in EMRs	Corresponding rate (%)
Body signs	5	4	80.00
Diseases history	23	14	60.86
Smoking and alcohol history	12	8	66.67
Exercise history	2	0	0.00
Medication history	4	4	100.00
Family diseases history	10	6	60.00
Amount	56	36	64.29

Results of data extraction from EMRs.

Data fields category	Amount of data fields	Amount of extraction rules	Amount of extracted data items	Error extraction	Accuracy (%)
Body signs	4	4	432	6	98.61
Diseases history	14	17	876	12	98.63
Smoking and alcohol history	8	10	453	10	97.79
Exercise history	0	0	0	0	0.00
Medication history	**4**	**11**	**214**	**11**	**94.86**
Family disease history	6	2	276	7	97.46
Amounts	36	44	2251	46	97.96

Verification results of each experiment.

Experiment	Data errors found	Accurate counts	Miss inspections	Consumption of time (h)
Clinician A	502	442	180	6.0
Clinician B	569	527	95	9.0
Clinician C	479	453	168	7.5
Automated approach	641	598	24	0.5

Average manual approach vs automatic approach.

Method name	Accuracy (%)	Recall (%)	Consumption of time (h)
Manual approach	92.03	76.17	7.5
Automated approach	93.29	96.14	0.5

Tian et al., 2019

Conclusion

- ✓ Automated data verification with data from paper-based CRF and EMRs was developed which reveal data quality problems and improve the data quality of registry data.

- ✓ Compared to the manual data verification approach, the automated approach has a higher recall of identified data errors and accuracy. At the same time, the time consumed is far less.

- ✓ Chinese Coronary Artery Diseases results suggest that the automated approach is more effective and efficient for identifying incomplete data and incorrect data in a registry.

- ✓ Further studies to enhance the accuracy and recall of this approach and correct some data errors automatically should be carried out.

References

1) List of Registries, Natl Inst Health NIH, 2015 (Accessed 31 January 2018). https://www.nih.gov/health-information/nih-clinical-research-trials-you/ list-registries.

2) Electronic Data Capture in Clinical Trials | Applied Clinical Trials n.d. http:// www.appliedclinicaltrialsonline.com/electronic-data-capture-clinical-trials (Accessed 11 February 2019).

3) MNIST handwritten digit database, Yann LeCun, Corinna Cortes and Chris Burges n.d. http://yann.lecun.com/exdb/mnist/ (Accessed 27 March 2018).

4) Ioffe S., Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift. 2015 ArXiv:150203167.

5) J. Lei, B. Tang, X. Lu, K. Gao, M. Jiang, H. Xu, A comprehensive study of named entity recognition in Chinese clinical text, J. Am. Med. Inform. Assoc. 21 (2014) 808–814, doi:10.1136/amiajnl-2013-002381.

6) H. Wang, W. Zhang, Q. Zeng, Z. Li, K. Feng, L. Liu, Extracting important information from Chinese Operation Notes with natural language processing methods, J. Biomed. Inform. 48 (2014) 130–136, doi:10.1016/j.jbi.2013.12.017.

7) Coronary Clinical Imaging Big Data Platform. https://www.ct-registry.cn/#/ login (Accessed 28 March 2018).

8) N. Sheetz, B. Wilson, J. Benedict, E. Huffman, A. Lawton, M. Travers, et al., Evaluating source data verification as a quality control measure in clinical trials, Ther. Innov. Regul. Sci. 48 (2014) 671–680, doi: 10.1177/2168479014554400 .

9) Risk-Based Monitoring in Clinical Trials: Getting Started n.d. https://www.jmp. com/en _nl/software/clinical- data- analysis- software/risk- based- monitoring. html (accessed 28 March 2018).

Thank you

The Global Healthcare Data Science Community

Contact Channels
UK +44 1843 609600
USA +1 609 514 5105
office@phuse.global
phuse.global

Social Media
@phusetwitta
/phusebook
/phusetube
/company/phuse