

The radiological report

Citation for published version (APA):

Nobel, J. M. (2023). *The radiological report: a compromise between structured reporting and natural language processing*. [Doctoral Thesis, Maastricht University]. Maastricht University. <https://doi.org/10.26481/dis.20230928jn>

Document status and date:

Published: 01/01/2023

DOI:

[10.26481/dis.20230928jn](https://doi.org/10.26481/dis.20230928jn)

Document Version:

Publisher's PDF, also known as Version of record

Please check the document version of this publication:

- A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
- The final author version and the galley proof are versions of the publication after peer review.
- The final published version features the final layout of the paper including the volume, issue and page numbers.

[Link to publication](#)

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license above, please follow below link for the End User Agreement:

www.umlib.nl/taverne-license

Take down policy

If you believe that this document breaches copyright please contact us at:

repository@maastrichtuniversity.nl

providing details and we will investigate your claim.



THE RADIOLOGICAL REPORT

A compromise between Structured Reporting
and Natural Language Processing

MARTIJN NOBEL

THE RADIOLOGICAL REPORT

*A compromise between Structured Reporting
and Natural Language Processing*

ISBN: 978-94-6473-177-4

Layout: J.M. Nobel

Cover design: by converting the title of this thesis: “The Radiological Report: A compromise between Structured Reporting and Natural Language Processing”, into an image using the AI-tool stable diffusion 2.1

<https://huggingface.co/spaces/stabilityai/stable-diffusion>

Printed and published by: proefschriften.nl

© J.M. Nobel

All rights reserved. No part of this publication may be reproduced, stored in a retrieval database or published in any form or by any means, electronic, mechanical or photocopying, recording or otherwise, without the prior written permission of the author, or, when appropriate, of the publishers of the publications.

THE RADIOLOGICAL REPORT

*A compromise between Structured Reporting
and Natural Language Processing*

Proefschrift

ter verkrijging van de graad van doctor aan de Universiteit Maastricht,
op gezag van Rector Magnificus, prof. dr. Pamela Habibović,
volgens het besluit van het College van Decanen,
in het openbaar te verdedigen
op 28 september 2023, om 10.00 uur

door

Johan Martijn Nobel

Promotores

Prof. dr. S.G.F. Robben

Prof. dr. ir. A.L.A.J. Dekker

Assessment Committee

Prof. dr. J.E. Wildberger (Chairman)

Prof. dr. J.C. Scholtes

Dr. W.B. Veldhuis

University Medical Center Utrecht

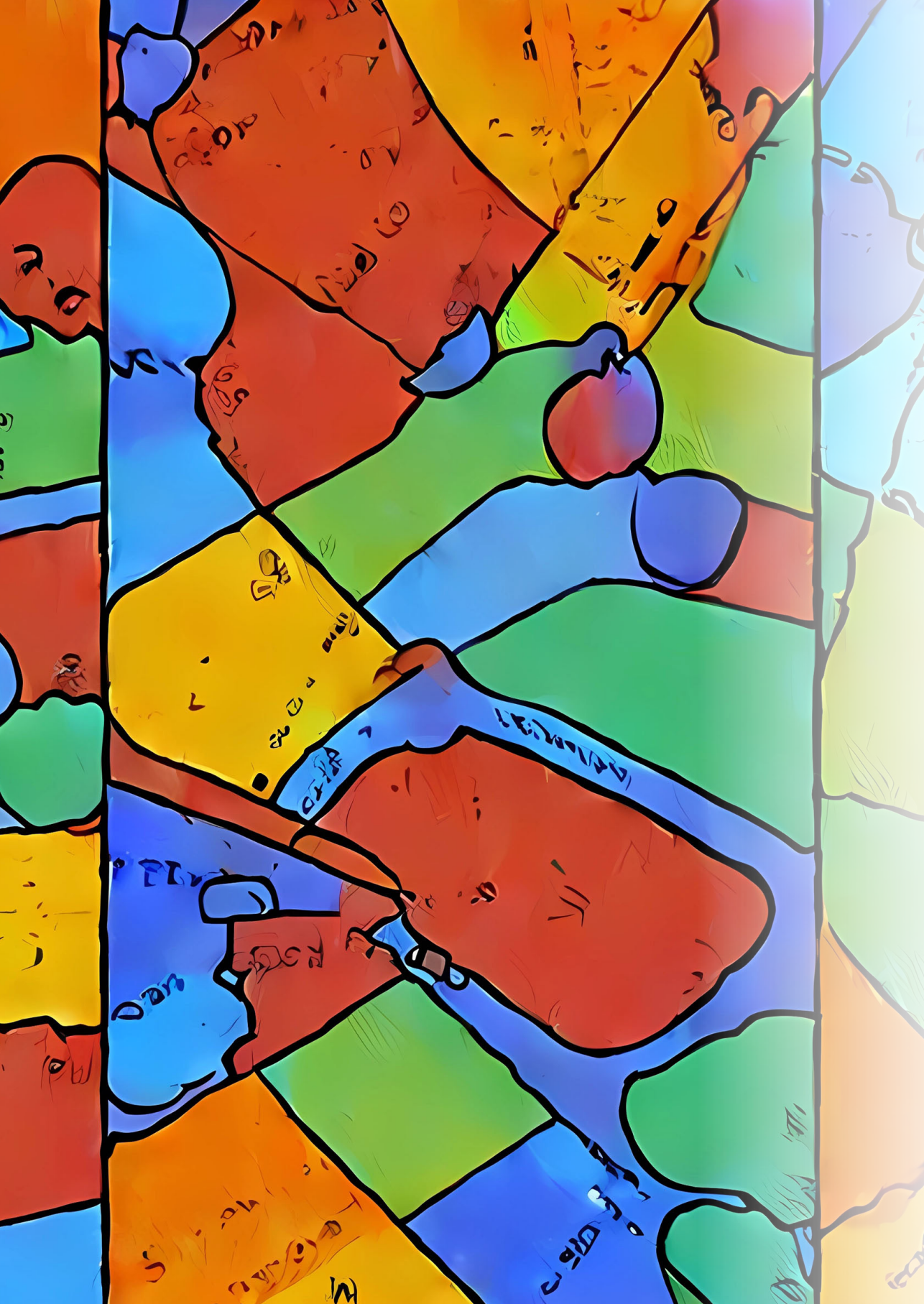
Dr. ir. P.M.A. van Ooijen

University Medical Center Groningen

Index

INTRODUCTION	11
Chapter 1: Introduction, aims and outline of thesis	13
PART ONE: STRUCTURED REPORTING IN RADIOLOGY	29
Chapter 2: Redefining the structure of structured reporting in radiology	31
Chapter 3: Structured reporting in radiology: a systematic review to explore its potential	43
PART TWO: NATURAL LANGUAGE PROCESSING	75
Chapter 4: Natural Language Processing in Dutch free text radiology reports: challenges in a small language area staging pulmonary oncology	77
Chapter 5: T-staging pulmonary oncology from radiological reports using natural language processing: translating into a multi-language setting	97
Chapter 6: Automated pulmonary oncology staging from free text radiological reports: extending the Dutch algorithm towards full utilization	123
Chapter 7: Natural Language Processing algorithm used for Staging Pulmonary Oncology from Free-text Radiological Reports: validation toward clinical use	147
GENERAL DISCUSSION AND SUMMARY	169
Chapter 8: General discussion	171
Chapter 9: Summary	193
ADDENDUM	197
Nederlandstalige samenvatting	198
Impact	200
Curriculum Vitae	203
List of publications	204
Dankwoord	206

INTRODUCTION





Chapter 1:

Introduction, aims and outline of thesis

General introduction

PART I: Radiological report and reporting process

The radiological report is the main output format of the radiologist to the referring clinician or general practitioner (GP) and is considered the golden standard in radiology communication [1-5]. In fact, it is a medicolegal document [6-8] in which every important aspect about a particular entity can (and should) be described by the reporter [1-5, 9-10]. Therefore, this document is very important in radiological practice, but also in the clinical process, as it is the translational step between the medical question and the interpretation of the findings on the radiological examination [11, 12]. The reporting process is complex, in which the answer of the medical question should be accurately presented in the radiological report in a readable fashion. Therefore, two main objectives in radiology reporting can be distinguished: **accurate content** and **a readable structure** (Fig. 1).

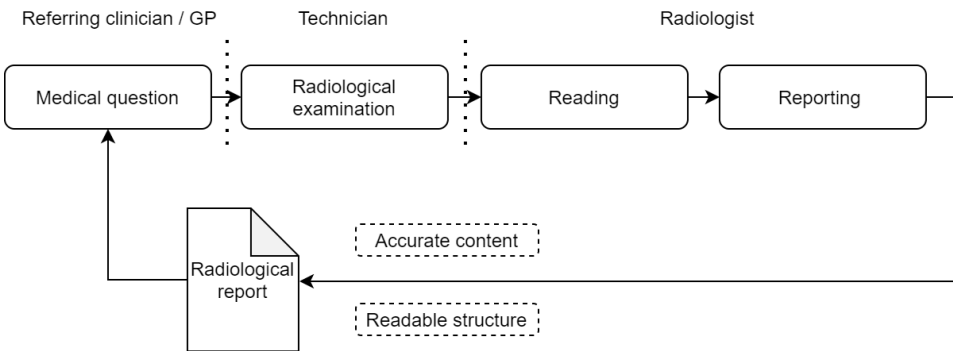


Figure 1. Radiological reporting process.

Digitalization in healthcare and the introduction of the Picture Archive and Communication System (PACS) as well as Radiology Information System (RIS) dramatically changed radiology practice [12-16]. This offered opportunities for changes in the way of reporting and for better access to the radiological report by the referring clinician and GP [2, 12]. The introduction of speech recognition in radiology has

enhanced the reporting process even more, as direct and faster reporting into the PACS was made possible [17-21].

However, since the beginning of reporting in radiology – somewhere close to the invention of the X-ray and the first radiological report in 1895 by Wilhelm Röntgen [22, 23] – little has changed in the reporting process. Of course, the radiological report is nowadays made with speech recognition software instead of a being a (hand)written document, but reporting principles and reporting style are still roughly the same [6, 10, 11, 24]. Actually, the radiological report still is commonly a free text document made by the radiologist.

Standardized reporting and structured reporting

Standardized reporting

Closely after the discovery of the X-ray, Hickey was one of the first to introduce some sort of streamlining into the radiological report [25]. Radiological reports at that time were ambiguous, and often the outcome of the examination and its report did not match the clinical condition [10, 23]. According to Hickey, standardization of the radiological report and the use of standardized nomenclature is key in improving radiology reports [25].

Despite the efforts put into standardization in order to streamline the content of the radiological report, the value of the radiological report up to this day is still very reporter-specific and probably education dependent [10, 11, 24, 26]. After all, there are radiologists that write short staccato reports and others that write large, prose (master)pieces. Therefore, the content or at least the style is radiologist-dependent [1, 27]. This wide variety in reporting manner leads to inconsistent reporting [1, 10, 26-29]. Langlotz states about this ongoing lack of report consistency [10]: “*Anyone who has attempted to glean definitive conclusions from even a small sample of radiology reports will agree we have a problem.*”

Especially the last decades, more attention and several attempts and guidelines on what to state in the radiological report have been published to increase report uniformity [1-

1

5, 24, 26-31]. Also many different standardization tools emerged to increase report uniformity using uniform language. Especially tools for risk assessment and quality assurance, such as Breast Imaging Reporting And Data System (BIRADS), Prostate Imaging Reporting And Data System (PIRADS), Thyroid Imaging And Data System (TIRADS) and for instance the Fleischner criteria on lung nodules have been created to help reporters describe important image findings, to do a risk estimation and to aid clinical decision making and follow up [32-35].

Structured reporting

More recently, the idea of improving reporting in radiology again gained interest when the term Structured Reporting (SR) emerged in literature as a possible solution for the need to accurately describe radiological findings as well as to implement reporting guidelines [10, 36-38]. Nowadays, the only rule set is that the radiological report should start with a clinical question, followed by the observations and should end with a section on findings [1, 27]. SR is supposed to aid the reporter by building the radiological report with a readable structure, by using for instance a strict format (template) or using an interactive report builder (drop-down menu). When doing so, the reporter is supported and guided to choose from different disease-specific options or locations out of a menu. Therefore, these items should not be reported any more. A different improvement of SR is that it helps the radiologist to report all items needed for the specific task with the right options and description. In literature, many advantages of its potential use have been described, as it can enhance and speed up workflow and increase accuracy and completeness. However, also several disadvantages are brought forward, as SR may hamper personal reporting freedom or burdening specific description of findings [39-42]. Still, large radiological societies, like the Radiological Society of North-America (RSNA) and the European Society of Radiology (ESR) combined forces to promote the use of SR [36, 43].

As such, both standardized reporting and structured reporting are supposed to increase the value of the radiological report (Fig. 2). Furthermore, in combination with already set or yet to be established reporting guidelines, it can streamline the radiological report.

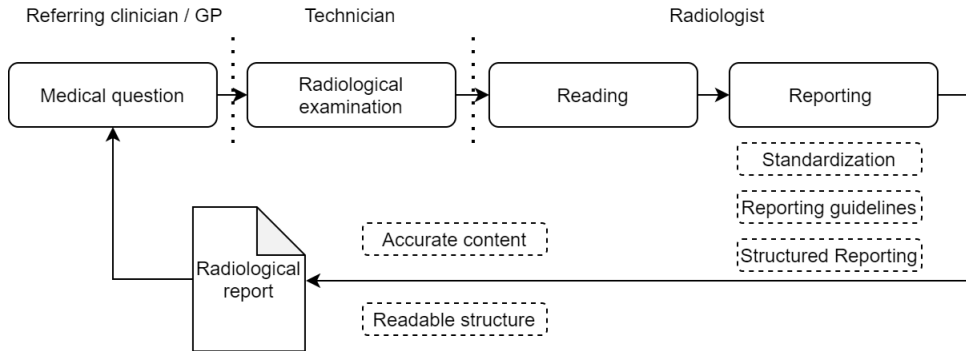


Figure 2. Radiological reporting process and tools to increase the value of the radiological report; standardization, reporting guidelines and structured reporting.

To allow for proper clinical implementation and adequate evaluation, the concepts of standardized reporting and SR need to be clear. However, various assumptions of what standardized reporting and SR involves are circulating in literature, which has led to confusion as to its actual meaning. Since the hype around SR started, relatively few studies define the term SR or have searched for evidence-based recommendations, whereas both may be pivotal for successful implementation. One of the omissions in this field is the lack of clear distinctions between both concepts and understanding of the differences between standardized reporting and structured reporting. As a consequence, it is unknown what the current status of SR in radiology reporting is and how it should be implemented best. Also despite the promotion by large radiological societies, it is necessary to know to what extent SR is being researched and implemented in clinical practice, to explore its level of evidence and to provide an overview of the current status of SR in radiology. Only then it is possible to find out whether SR actually enhances the radiological reporting process.

PART II: Text mining and Natural Language Processing (NLP)

Text mining

The aforementioned structured reporting mainly focuses on human-based interventions that can be supported by IT solutions in order to enhance the radiological

1

report. However, as artificial intelligence (AI) is becoming more accepted in the modern world, it is interesting to explore how AI can assist in creating and improving the radiological report.

Because of the process of digitalization in healthcare, very large quantities of patient data – patient follow ups, blood results as well as their medical history – are digitally stored in the medical Electronic Health Record (EHR) [44, 45]. Likewise, all radiological and nuclear medicine reports are being stored digitally in the EHR. This enormous data storage of medical information can be potentially used for many tasks, such as workflow improvement, quality assurance, education and research. However, most of this medical information is left unused, as it is stored in an unstructured manner and, as data retrieval of unstructured data is very laborious, it is not easy to (re)use [44-46]. Again, this is also true for data of the radiological report, which is mostly written as unstructured free text and therefore not easy to reuse.

Data mining [47, 48] is considered to be a solution for the extraction and search for specific data or correlations. In data mining, computing power is used to mine or search an enormous amount of unstructured data in order to find the appropriate data. By doing so, large quantities of data can be searched, processed, stored and labeled for all kinds of purposes, and without or with only little human interference. Text mining is a subtype of data mining and can be used for searching text files as for instance the radiological report. As image mining and radiomics are booming in radiology [49-51], text mining can be the next gamechanger, as it can process the free text radiological report and help improve the reporting process and the radiological report itself. In this context, text mining can be a method to structure unstructured free text data, thus functioning as a counterpart for SR. After all, in SR, text is inserted and stored in a structured way, but text mining can also assure this structured storage using computing power without the help of the reporter. By doing so, text mining can facilitate all kinds of postprocessing processes leading to, for instance, a structured report or specific relevant additions.

Natural Language Processing

Natural Language Processing (NLP) is an AI approach that facilitates understanding of human language by computer interpretation [44, 52-54]. It is a tool that can be used for text mining purposes including the radiological report [46, 54, 55]. The process of analyzing a free text radiological report using NLP is typically divided into two steps: 1) preprocessing; document cleaning and preparing and 2) processing; actual task execution. In the preprocessing step the report is sectionized, spelling is corrected, abbreviations are expanded into full text, sentences are being split and negations are checked. Also important concepts (words or word combinations), measurements and context are extracted in this step [56-57]. In the processing step, a particular task is executed and the report is processed resulting in a specific output depending of the task. Hence, the report content can be staged, annotated or specific information can be extracted or added as necessary for its specific goal. An important remark is that the amount of data and the difficulty level of the task are both factors that force to use a rule based, hybrid or machine learning approach [52, 54]. For instance, all approaches can be used with large quantities of data. However, when only a small amount of data is available, a rule based or hybrid approach can be used only, as machine learning is typically infeasible in a small data set.

Nowadays, NLP in healthcare is mainly applied in research settings, but there is already some routine use in radiology reporting [54, 56]. Different tasks in which NLP applications are used to mine the radiological report are cohort building, query-based case retrieval, clinical support, diagnostic support and quality assessment [54, 56]. Radiological applications that can be used in clinical practice mainly focus on diagnostic surveillance and clinical support. An example of a clinical supporting system is a tool that can detect the description for the term 'fracture' in the free text report, and, based on the type of described fracture, a real-time recommendation for an additional X-ray or MRI can be made [58]. The same is true for detection of pneumonia, in which a recommendation for antibiotics can be made based on the description in the radiological report [59, 60]. An example of a diagnostic surveillance tool is one that alerts a referring clinician in case an important outcome, such as thromboembolic disease, acute appendicitis or pneumonia, is being diagnosed [61-64]. The main goal of

all these NLP applications is to extract and process data, and thereby adding value to the radiological report (Fig. 3).

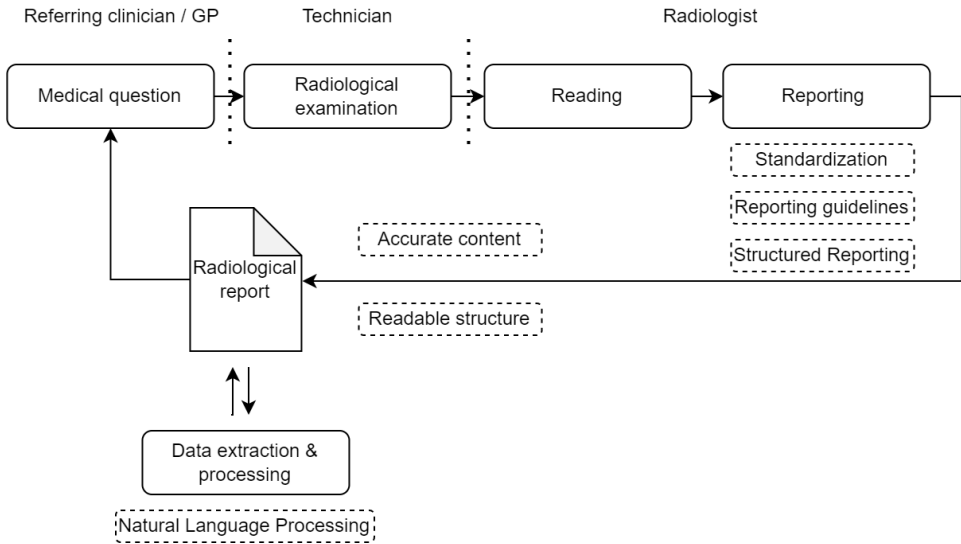


Figure 3. Radiology reporting process and tools to increase the value of the radiological report; standardization, reporting guidelines, structured reporting and Natural Language Processing via data extraction and processing.

NLP in oncology reporting

When focusing on the oncological setting, NLP can also be used to ascertain oncological outcomes for regular follow up outcomes, follow up of acute oncological findings, tumor recurrence rates or for cancer registries and oncological classification [52, 65-72]. However, NLP for TNM staging has not been widely used.

The oncological TNM classification system is a worldwide used and accepted classification system that assists with tumor-staging in oncological patients and stands for Tumor, (lymph)Node and Metastasis which are the important parameters in cancer staging [73]. Based on clinical, pathological and diagnostic information, an oncological patient is staged according to the TNM classification and adequate treatment can be given to achieve patients' best outcome.

Lung cancer is the most common oncological cause of death. Imaging is an important part of the diagnostic and staging process in lung cancer [74]. Each patient with this diagnosis will have a (PET-)CT of the chest to evaluate their cancer stage. The radiological report is used for communication of lung carcinoma staging and it is important that at least all items are described that are mandatory for tumor staging as mentioned in the TNM classification. In (PET-)CT reports, the description of tumor size, the local extension and tumor spread together determine a particular tumor stage. To assure this staging task, and to be as accurate and complete as possible, an NLP application might be used as a reporting support system.

It is the right time and very important to explore how NLP can be used in radiology for free text mining purposes, as it can enhance the radiological process, and the radiological report in particular. Especially the oncological staging process of, for instance, lung cancer is of interest because of its importance for patients' clinical staging and their treatment. Because this staging process is highly complex, it is a perfect use case to highlight the potential of such an NLP tool in radiology reporting. Probably even more importantly, it helps to explore its possibilities and find solutions for its imperfections.

Aims and outline of this thesis

The overall aim of this thesis is to better understand how to improve reporting in radiology. The projects that have led to this thesis focus on structured reporting and explore the usage of free text mining and NLP in radiology reporting.

The specific research aims are:

- To explore what structured reporting entails and what its definition is
- To summarize efforts done on the subject of structured reporting and whether structured reporting is evidence based
- To assess how free text mining and Natural Language Processing (NLP) can be used in a Dutch clinical setting concerning primary lung carcinoma T-staging according to the TNM classification system

- To assess how the rule-based NLP primary lung carcinoma T-staging algorithm can be translated and trained in an English setting
- To extend the existing Dutch NLP T-staging algorithm towards a TN-staging algorithm
- To extend the existing Dutch NLP TN-staging algorithm with PET-CT functionality and external validation

This thesis is divided into two parts in which the first part focuses on structured reporting and the second part on how to use free text mining and NLP in radiology reporting.

Chapter 2 elaborates on the current interpretation of what Structured Reporting is and suggests a proper definition of standardization and structured reporting.

Chapter 3 is a narrative systematic review in which the evidence for structured reporting is assessed and gives an overview on the clinical implementation and outcomes.

In **Chapter 4**, free text Dutch radiology reports are used to train and validate a rule based free text NLP algorithm that is capable of T-staging primary lung carcinoma according to the TNM oncology staging system.

Chapter 5 describes the process of translating, training and validating the Dutch free text NLP T-staging algorithm for staging primary lung carcinoma into English, to explore its functionality in a different language.

In **Chapter 6**, the extension of the Dutch NLP T-staging algorithm capable of staging primary lung carcinoma towards a TN-staging algorithm is described.

Chapter 7 elaborates on adding an extra PET-CT functionality layer upon the already existing Dutch NLP TN-staging algorithm that can be used for staging lung carcinoma according to the TNM classification system.

The current situation and knowledge as well as future perspectives regarding structured reporting and free text mining in radiology are discussed in **Chapter 8**.

References

1. European Society of Radiology (ESR). Good practice for radiological reporting. Guidelines from the European Society of Radiology (ESR). *Insights Imaging*. 2011;2(2):93-96. doi:10.1007/s13244-011-0066-7.
2. Grieve FM, Plumb AA, Khan SH. Radiology reporting: a general practitioner's perspective. *Br J Radiol*. 2010;83(985):17-22. doi: 10.1259/bjr/16360063.
3. Recommandations générales pour l'élaboration d'un compte-rendu radiologique (CRR). *J Radiol*. 2007;88(2):304-6. doi: 10.1016/S0221-0363(07)89822-2.
4. American College of Radiology. ACR practice guideline for communication of diagnostic imaging findings [Internet]. Reston: American College of Radiology; 2005 [cited September 2020]. Available from <https://www.acr.org/-/media/acr/files/practice-parameters/communicationdiag.pdf>
5. The Royal College of Radiologists. Standards for the Reporting and Interpretation of Imaging Investigations [Internet]. London: The Royal College of Radiologists; 2006 [cited September 2020]. Available from https://www.rcr.ac.uk/sites/default/files/bfcro61_standardsforreporting.pdf
6. Wallis A, McCoubrie P. The radiology report - are we getting the message across? *Clin Radiol*. 2011;66(11):1015-22. doi: 10.1016/j.crad.2011.05.013.
7. Berlin L. Pitfalls of the vague radiology report. *AJR Am J Roentgenol*. 2000;174(6):1511-8. doi: 10.2214/ajr.174.6.1741511.
8. Eisenberg RL. *Radiology and the Law: Malpractice and Other Issues*. New York: Springer; 2003.
9. Siström CL, Langlotz CP. A framework for improving radiology reporting. *J Am Coll Radiol*. 2005;2:159e67. doi: 10.1016/j.jacr.2004.06.015.
10. Langlotz CP. *The radiology report: a guide to thoughtful communication for radiologists and other medical professionals*. CreateSpace Independent Publishing Platform; 2015.
11. Brady AP. Radiology reporting - from Hemingway to HAL? *Insights Imaging*. 2018;9(2):237-246. doi:10.1007/s13244-018-0596-3.
12. Weiss DL, Kim W, Branstetter IV BF, Prevedello LM. Radiology reporting: a closed-loop cycle from order entry to results communication. *J Am Coll Radiol*. 2014;11(12):1226-37. doi: 10.1016/j.jacr.2014.09.009.
13. Weiss DL, Kim W, Branstetter IV BF, Prevedello LM. Radiology reporting: a closed-loop cycle from order entry to results communication. *J Am Coll Radiol*. 2014;11(12):1226-37. doi: 10.1016/j.jacr.2014.09.009.
14. Joshi V, Narra VR, Joshi K, Lee K, Melson D. PACS administrators' and radiologists' perspective on the importance of features for PACS selection. *J Digit Imaging*. 2014;27(4):486-95. doi: 10.1007/s10278-014-9682-3.
15. Geis JR. Medical imaging informatics: how it improves radiology practice today. *J Digit Imaging*. 2007;20(2):99-104. doi: 10.1007/s10278-007-9010-2.

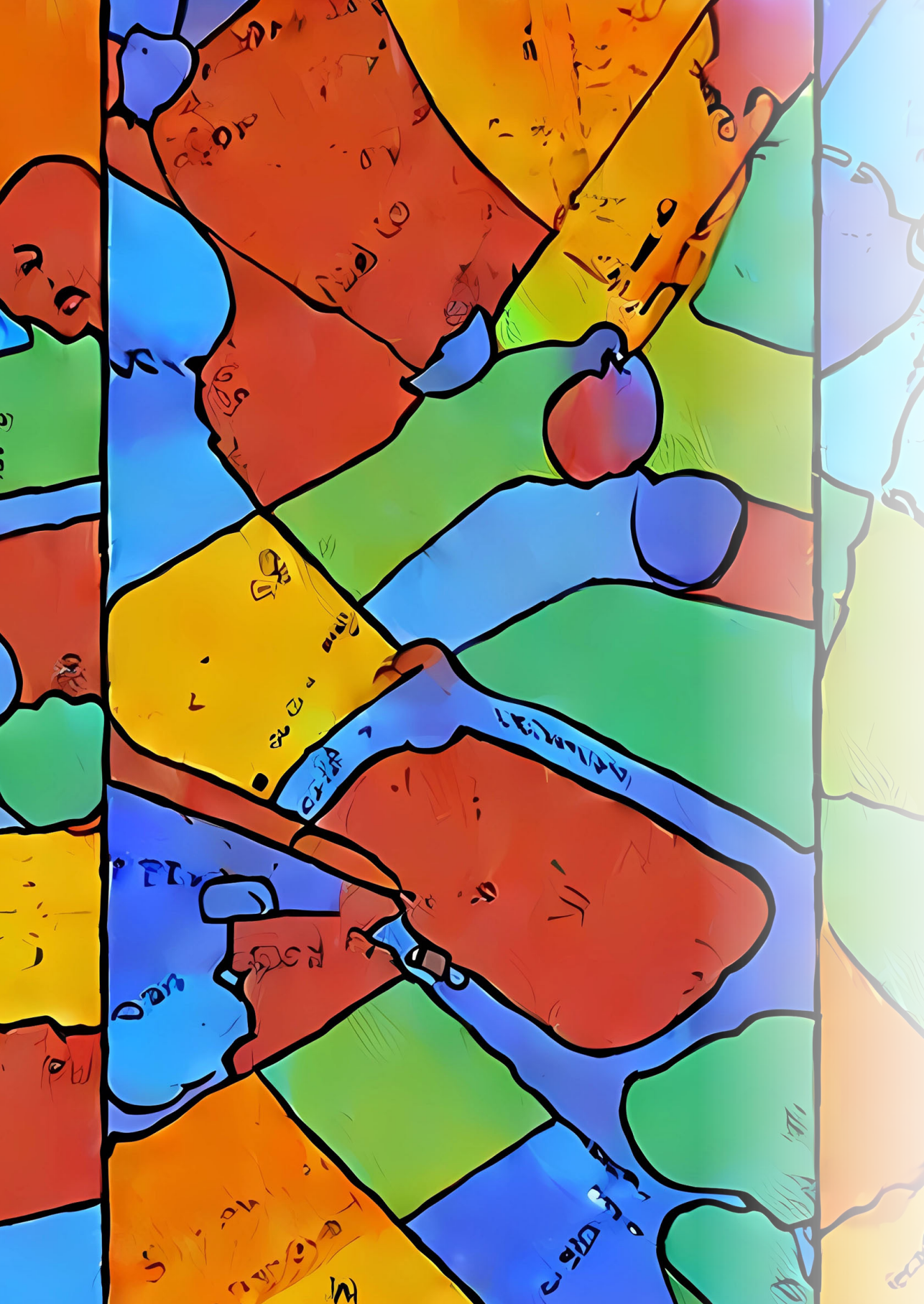
16. Weiss DL, Bolos PR. Reporting and dictation. In: Branstetter IV BF, editor. Practical imaging informatics: foundations and applications for PACS professionals. New York: Springer; 2009. p. 147-162.
17. Creighton C. A literature review on communication between picture archiving and communication systems and radiology information systems and/or hospital information systems. *J Digit Imaging.* 1999;12(3):138-43. doi: 10.1007/BF03168632.
18. Liu D, Zucherman M, Tulloss Jr. WB. Six characteristics of effective structured reporting and the inevitable integration with speech recognition. *J Digit Imaging.* 2006;19:98-104. doi: 10.1007/s10278-005-8734-0.
19. Glaser C, Trumm C, Nissen-Meyer S, Francke M, Küttner B, Reiser M. Spracherkennung: Auswirkung auf Workflow und Befundverfügbarkeit [Speech recognition: impact on workflow and report availability]. *Radiologe.* 2005;45(8):735-42. doi: 10.1007/s00117-005-1253-7.
20. Kauppinen T, Koivikko MP, Ahovuo J. Improvement of report workflow and productivity using speech recognition - a follow-up study. *J Digit Imaging.* 2008;21(4):378-82. doi: 10.1007/s10278-008-9121-4. Erratum in: *J Digit Imaging.* 2008;21(4):383.
21. Reiner BI. Expanding the functionality of speech recognition in radiology: creating a real-time methodology for measurement and analysis of occupational stress and fatigue. *J Digit Imaging.* 2013;26(1):5-9. doi: 10.1007/s10278-012-9540-0.
22. Zonneveld FW. Spectacular rediscovery of the original prints of radiographs Roentgen sent to Lorentz in 1896. *Insights Imaging.* 2020;11(1):46. doi: 10.1186/s13244-020-00846-x.
23. Röntgen WC. Eine Neue Art von Strahlen. Würzburg (Germany): Medicophysical Institute of the University of Würzburg; 1896.
24. Reiner BI, Knight N, Siegel EL. Radiology reporting, past, present, and future: the radiologist's perspective. *J Am Coll Radiol.* 2007;4(5):313-9. doi: 10.1016/j.jacr.2007.01.015.
25. Hickey PM. Standardization of roentgen-ray reports. *AJR Am J Roentgenol.* 1922;9:442-445.
26. European Society of Radiology (ESR). ESR concept paper on value-based radiology. *Insights Imaging.* 2017;8(5):447-454. doi: 10.1007/s13244-017-0566-1.
27. Kahn CE Jr, Langlotz CP, Burnside ES, Carrino JA, Channin DS, Hovsepian DM, et al. Toward best practices in radiology reporting. *Radiology.* 2009;252(3):852-6. doi: 10.1148/radiol.2523081992.
28. Lukaszewicz A, Uricchio J, Gerasymchuk G. The Art of the Radiology Report: Practical and Stylistic Guidelines for Perfecting the Conveyance of Imaging Findings. *Can Assoc Radiol J.* 2016;67(4):318-321. doi: 10.1016/j.carj.2016.03.001.
29. Reiner BI. The challenges, opportunities, and imperative of structured reporting in medical imaging. *J Digit Imaging.* 2009;22(6):562-8. doi: 10.1007/s10278-009-9239-z.
30. Hall FM. Language of the radiology report: primer for residents and wayward radiologists. *AJR Am J Roentgenol.* 2000;175(5):1239-42. doi: 10.2214/ajr.175.5.1751239.
31. Jacoby J, Ayer R, editors. Frameworks for radiology reporting. London: Taylor and Francis; 2009.

32. D'Orsi CJ, Sickles EA, Mendelson EB, Morris EA. ACR BI-RADS® Atlas: Breast Imaging Reporting and Data System. Reston (VA): American College of Radiology; 2013.
33. Turkbey B, Choyke PL. PIRADS 2.0: what is new? *Diagn Interv Radiol.* 2015;21(5):382–384
34. Tessler FN, Middleton WD, Grant EG, Hoang JK, Berland LL, Teefey SA, et al. ACR Thyroid Imaging, Reporting and Data System (TI-RADS): White Paper of the ACR TI-RADS Committee. *J Am Coll Radiol.* 2017;14(5):587–595. doi: 10.1016/j.jacr.2017.01.046.
35. MacMahon H, Naidich DP, Goo JM, Lee KS, Leung ANC, Mayo JR, et al. Guidelines for Management of Incidental Pulmonary Nodules Detected on CT Images: From the Fleischner Society 2017. *Radiology.* 2017;284(1):228–243. doi: 10.1148/radiol.2017161659.
36. European Society of Radiology (ESR). ESR paper on structured reporting in radiology. *Insights Imaging.* 2018;9(1):1–7. doi: 10.1007/s13244-017-0588-8.
37. Powell DK, Silberzweig JE. State of Structured Reporting in Radiology, a Survey. *Acad Radiol.* 2015;22:226–33. doi: 10.1016/j.acra.2014.08.014.
38. Radiological Society of North America. RadReport template library [Internet]. Oak Brook (IL): Radiological Society of North America; 2020 [cited 15 Dec 2020] Available from <https://radreport.org>
39. Bosmans JM, Peremans L, Menni M, De Schepper AM, Duyck PO, Parizel PM. Structured reporting: if, why, when, how-and at what expense? Results of a focus group meeting of radiology professionals from eight countries. *Insights Imaging.* 2012;3(3):295–302. doi: 10.1007/s13244-012-0148-1.
40. Haroun RR, Al-Hihi MM, Abujudeh HH. The Pros and Cons of Structured Reports. *Curr Radiol Rep.* 2019;7:31. doi: 10.1007/s40134-019-0342-8.
41. Ganeshan D, Duong PT, Probyn L, Lenchik L, McArthur TA, Retrouvey M, et al. Structured Reporting in Radiology. *Acad Radiol.* 2018;25(1):66–73. doi: 10.1016/j.acra.2017.08.005.
42. Weiss DL, Langlotz CP. Structured reporting: patient care enhancement or productivity nightmare? *Radiology.* 2008;249(3):739–47. doi: 10.1148/radiol.2493080988.
43. Radiological Society of North America. RadReport template library [Internet]. Oak Brook (IL): Radiological Society of North America; 2020 [cited 15 Dec 2020] Available from <https://radreport.org>
44. Kreimeyer K, Foster M, Pandey A, Arya N, Halford G, Jones SF, et al. Natural language processing systems for capturing and standardizing unstructured clinical information: A systematic review. *J Biomed Inform.* 2017;73:14–29. doi: 10.1016/j.jbi.2017.07.012.
45. Pinto Dos Santos D, Baessler B. Big data, artificial intelligence, and structured reporting. *Eur Radiol Exp.* 2018;2(1):42. doi: 10.1186/s41747-018-0071-4.
46. Mozayan A, Fabbri AR, Maneveese M, Tocino I, Chheang S. Practical Guide to Natural Language Processing for Radiology. *Radiographics.* 2021;41(5):1446–1453. doi: 10.1148/rg.2021200113.
47. Smyth P. Data mining: data analysis on a grand scale? *Stat Methods Med Res.* 2000;9(4):309–27. doi: 10.1177/096228020000900402.
48. Lambin P, Rios-Velazquez E, Leijenaar R, Carvalho S, van Stiphout RG, Granton P, et al. Radiomics: extracting more information from medical images using advanced feature analysis. *Eur J Cancer.* 2012;48(4):441–6. doi: 10.1016/j.ejca.2011.11.036.

49. Lambin P, Leijenaar RTH, Deist TM, Peerlings J, de Jong EEC, van Timmeren J, et al. Radiomics: the bridge between medical imaging and personalized medicine. *Nat Rev Clin Oncol*. 2017;14(12):749-762. doi: 10.1038/nrclinonc.2017.141.
50. Guiot J, Vaidyanathan A, Deprez L, Zerka F, Danthine D, Frix AN, et al. A review in radiomics: Making personalized medicine a reality via routine imaging. *Med Res Rev*. 2022;42(1):426-440. doi: 10.1002/med.21846.
51. Reginelli A, Nardone V, Giacobbe G, Belfiore MP, Grassi R, Schettino F, et al. Radiomics as a New Frontier of Imaging for Cancer Prognosis: A Narrative Review. *Diagnostics*. 2021;11(10):1796. doi: 10.3390/diagnostics11101796.
52. Yim WW, Yetisgen M, Harris WP, Kwan SW. Natural Language Processing in Oncology: A Review. *JAMA Oncol*. 2016;2(6):797-804. doi: 10.1001/jamaoncol.2016.0213.
53. Cai T, Giannopoulos AA, Yu S, Kelil T, Ripley B, Kumamaru KK, et al. Natural Language Processing Technologies in Radiology Research and Clinical Applications. *Radiographics*. 2016;36(1):176-91. doi: 10.1148/rg.2016150080.
54. Pons E, Braun LMM, Hunink MGM, Kors JA: Natural language processing in radiology: A systematic review. *Radiology*. 2016;279:329- 343. doi: 10.1148/radiol.16142770.
55. Cáceres SB. Electronic health records: beyond the digitization of medical files. *Clinics*. 2013;68(8):1077-1078. doi:10.6061/clinics/2013(08)02.
56. Sorin V, Barash Y, Konen E, Klang E. Deep Learning for Natural Language Processing in Radiology- Fundamentals and a Systematic Review. *J Am Coll Radiol*. 2020;17(5):639-648. doi: 10.1016/j.jacr.2019.12.026.
57. Nadkarni PM, Ohno-Machado L, Chapman WW. Natural language processing: an introduction. *J Am Med Inform Assoc*. 2011;18(5):544-51. doi: 10.1136/amiajnl-2011-000464.
58. Do BH, Wu AS, Maley J, Biswal S. Automatic retrieval of bone fracture knowledge using natural language processing. *J Digit Imaging*. 2013;26(4):709-713. doi: 10.1007/s10278-012-9531-1.
59. Fiszman M, Chapman WW, Aronsky D, Evans RS, Haug PJ. Automatic detection of acute bacterial pneumonia from chest x-ray reports. *J Am Med Inform Assoc*. 2000;7(6):593-604. doi: 10.1136/jamia.2000.0070593.
60. Fiszman M, Chapman WW, Evans SR, Haug PJ. Automatic identification of pneumonia related concepts on chest x-ray reports. *Proc AMIA Symp*. 1999:7-71.
61. Pham AD, Névéol A, Lavergne T, et al. Natural language processing of radiology reports for the detection of thromboembolic diseases and clinically relevant incidental findings. *BMC Bioinformatics*. 2014;15:266. doi: 10.1186/1471-2105-15-266.
62. Rink B, Roberts K, Harabagiu S, et al. Extracting actionable findings of appendicitis from radiology reports using natural language processing. *AMIA Jt Summits Transl Sci Proc*. 2013;2013:221.
63. Mendonça EA, Haas J, Shagina L, Larson E, Friedman C. Extracting information on pneumonia in infants using natural language processing of radiology reports. *J Biomed Inform*. 2005;38(4):314-321. doi: 10.1016/j.jbi.2005.02.003.

64. Haas JP, Mendonça EA, Ross B, Friedman C, Larson E. Use of computerized surveillance to detect nosocomial pneumonia in neonatal intensive care unit patients. *Am J Infect Control* 2005;33(8):439-443. doi: 10.1016/j.ajic.2005.06.008.
65. Lee SJ, Weinberg BD, Gore A, Banerjee I. A Scalable Natural Language Processing for Inferring BT-RADS Categorization from Unstructured Brain Magnetic Resonance Reports. *J Digit Imaging*. 2020;33(6):1393-1400. doi: 10.1007/s10278-020-00350-0.
66. Zeng Z, Espino S, Roy A, Li X, Khan SA, Clare SE, et al. Using natural language processing and machine learning to identify breast cancer local recurrence. *BMC Bioinformatics*. 2018;19(Suppl 17):498. doi: 10.1186/s12859-018-2466-x.
67. Lou R, Lalevic D, Chambers C, Zafar HM, Cook TS. Automated detection of radiology reports that require follow-up imaging using natural language processing feature engineering and machine learning classification. *J Digit Imaging*. 2020;33(1):131-136. doi: 10.1007/s10278-019-00271-7.
68. Abdulsalam AKA, Garvin JH, Redd A, Carter ME, Sweeny C, Meystre SM. Automated extraction and classification of cancer stage mentions from unstructured text fields in a central cancer registry. *AMIA Jt Summits Transl Sci Proc*. 2018;2017:16-25.
69. Kehl KL, Elmarakeby H, Nishino M, Van Allen EM, Lepisto EM, Hassett MJ, et al. Assessment of Deep Natural Language Processing in Ascertaining Oncologic Outcomes From Radiology Reports. *JAMA Oncol*. 2019;5(10):1421-1429. doi: 10.1001/jamaoncol.2019.1800.
70. Cheng LT, Zheng J, Savova GK, Erickson BJ. Discerning tumor status from unstructured MRI reports: completeness of information in existing reports and utility of automated natural language processing. *J Digit Imaging* 2010;23(2):119-132. doi: 10.1007/s10278-009-9215-7.
71. Jain NL, Friedman C. Identification of findings suspicious for breast cancer based on natural language processing of mammogram reports. *Proc AMIA Annu Fall Symp*. 1997:829-833.
72. Percha B, Nassif H, Lipson J, Burnside E, Rubin D. Automatic classification of mammography reports by BI-RADS breast tissue composition class. *J Am Med Inform Assoc*. 2012;19(5):913-916. doi: 10.1136/amiajnl-2011-000607.
73. Brierley J, Gospodarowicz MK, Wittekind C, editors. *TNM classification of malignant tumours*. 8th ed. Chichester: John Wiley & Sons Inc; 2017.
74. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J Clin*. 2021;71(3):209-249. doi: 10.3322/caac.21660.

PART ONE:
STRUCTURED REPORTING
IN RADIOLOGY





Chapter 2:

Redefining the structure of structured reporting in radiology

J. Martijn Nobel, Ellen M. Kok, Simon G.F. Robben

Insights into Imaging (2020)

Abstract

Structured reporting is advocated as a means of improving reporting in radiology to the ultimate benefit of both radiological and clinical practice. Several large initiatives are currently evaluating its potential. However, with numerous characterizations of the term in circulation, ‘structured reporting’ has become ambiguous and is often confused with ‘standardization’, which may hamper proper evaluation and implementation in clinical practice. This paper provides an overview of interpretations of structured reporting and proposes a clear definition that differentiates structured reporting from standardization. Only a clear uniform definition facilitates evidence-based implementation, enables evaluation of its separate components, and supports (meta-) analyses of literature reports.

Background

Structured reporting is a buzzword in radiology used to refer to a potential means of improving the quality of radiology reports [1]. In their statement paper, the European Society of Radiology (ESR) state that quality, datafication/quantification and accessibility are the main functional needs for moving from traditional free text reporting to standardized and structured reporting [2].

Structured reporting is thought to improve consistency and reproducibility of the radiological report. This improves readability and clarity of the radiological report, but also facilitates data mining in clinical or research settings.

Introduction of structured reporting led to the launch of several initiatives in the field and to numerous publications [3]. The main purpose of most published articles has been to describe the process of improving radiological reports by implementing “structured reporting”. However, various assumptions of what structured reporting involves are circulating in the literature, which has led to confusion as to its actual meaning. Relatively few studies define the term structured reporting or search for evidence-based recommendations, whereas both may be pivotal to successful implementation. This paper aims to redefine structured reporting by proposing distinctions between standardization and structured reporting.

Definition

In the statement paper on structured reporting in radiology the ESR makes a valuable contribution to the understanding of structured reporting and its implementation [3]. The society clearly describes the necessity of structured reporting in clinical practice by addressing a) the requirements and b) implementation strategies. They state that “the need to use uniform language and structure to accurately discuss findings in radiology is the basis for developing the concept of structured reporting” [3]. In their statement paper, a definition for structured reporting is set by describing three levels of structured reporting according to Weiss [4]:

1. Structured format: which paragraph(s) or subheading(s) should be used?
2. Consistent organization: which items should be reported in which order?
3. Consistent use of dedicated terminology: which lexicon or ontology should be implemented (i.e. standard language)?

This definition describes levels of structured reporting but does not address the definition of structured reporting itself. Actually, these levels address both standardized reporting as well as structured reporting, but do not highlight its separate function. We agree that there is a need for standardization: standardization of the format of the report, standardization of the medical content and standardization of vocabulary used. However, standardized reporting is not the same as structured reporting. Thus, an important step towards a uniform definition is to differentiate between standardization and structured reporting.

What is standardization?

Almost 100 years ago, Hickey suggested standardization in X-ray reporting, stating that it should “streamline [the] reporting manner and nomenclature to increase the value of the written report and its scientific accuracy” [5]. This definition is still relevant today, because standardization is aimed at improving the accuracy of the medical content of a radiological report.

Investigations in this field have focused on whether the radiological report can match a certain standard, such as content or lay-out, or whether using certain unambiguous vocabulary is feasible. Grading systems such as BI-RADS [6], PI-RADS [7], and lexicons

such as RadLex [8], are initiatives developed to increase the level of standardization in the radiological report. Such initiatives are considered to streamline and enhance understanding of the medical content, thus improving accuracy.

Proposed definition: Standardized reporting is a means of streamlining the medical content of a radiological report

What is (real) structured reporting?

Unlike standardized reporting, the definition of structured reporting is less clear in current literature. There is a wide variety of definitions, which makes the subject difficult to investigate and implement. Three recent examples are:

a) “A report is qualified as structured when all of the relevant information and diagnostic impressions are included, following specific terms and descriptors previously defined, as well as a predefined design” [9]

b) “Structured reporting is “the creation of standardized, organized information from templates via menus into a natural-sounding language report” [10]

c) “Structured reporting means the use of predefined formats and terms to create reports; in this sense, structured reports are those based on templates or checklists” [11]

A common factor seen in most definitions is that structured reporting must help the writer create their report, through either a predefined design, template, or a checklist. In 2005, Siström et al. [12] stated that “structured reporting represents simply one set of computer tools aimed at reducing variability and enhancing the clinical utility of formal radiology interpretations.” This adds to previous definitions in that structured reporting should be a computer tool that helps the reporter generate the report. To our mind, this is the clue to understanding the term structured reporting.

Proposed definition: Structured reporting is the use of an IT-based means of importing and arranging medical content in the radiological report.

In addition to the definition of the ESR, we pose that structured reporting is the way of creating the actual report by means of IT. By creating this distinction it is possible to appreciate two independent factors which independently can influence the report quality. One being standardization and one being the way of creating the report.

Our definition of structured reporting is more similar to the definition as proposed by Weiss et al. [13]. They distinguish between the use of templates or macros ('level 1') and structured reporting ('level 2'): a template or macro is a blueprint for the definitive report, and structured reporting is the tool used to convert medical content into the report. We propose to name level 1 structured lay-out [10], and level 2 structured content.

Level 1: Structured lay-out

Structured lay-out presents the findings in a strict, predefined order, creating and maintaining uniformity. It looks like a template or blueprint of the report. For example, standard headers such as title of examination, history/indication, technique, comparison, findings and conclusion create consistency [14]. In addition, standard sections can be used to indicate content, and subdivisions can be used to arrange longer reports. Examples include 'head to toe', 'hierarchical', which implies that the most important items are reported first, or 'itemized', in which a fixed ordering such as 'heart-lungs-liver-spleen-pancreas-etc.' is used [13, 15] (Fig. 1).

<p>Standardized (head to toe) Multinodular struma. A consolidation in the left upper lobe. Ruptured abdominal aneurysm with free retroperitoneal fluid.</p>	<p>Itemized</p> <p>Heart: normal dimensions Lungs: consolidation left upper lobe Liver: normal Spleen: normal Pancreas: normal Kidneys: normal Aorta: ruptured abdominal aneurysm with free retroperitoneal fluid Lymph nodes: none enlarged Thyroid: multinodular struma</p>
<p>Hierarchical Ruptured abdominal aneurysm with free retroperitoneal fluid. A consolidation in the left upper lobe. Multinodular struma.</p>	

Figure 1. Structured reporting level 1: Structured lay-out. Examples of structured lay-out. Standardized reports use a standardized order (free text), in hierarchical reports the most important items are mentioned first. The itemized report uses fixed headings.

Level 2: Structured content

Structured content is the manner in which the medical content is arranged and displayed in the report. This is the more technical aspect of IT-guided content generation. Examples mentioned in literature are drop down menus [16], pick lists [17] or point-and-click systems [12, 18, 19]. Gap filling is another form of structured content reporting, where blanks left in sentences must be filled with a specific phrase or word. One example of this concept is flowchart-guided input, such as SPIDER (Structured Platform-Independent Data Entry and Reporting) [18] (Fig. 2).

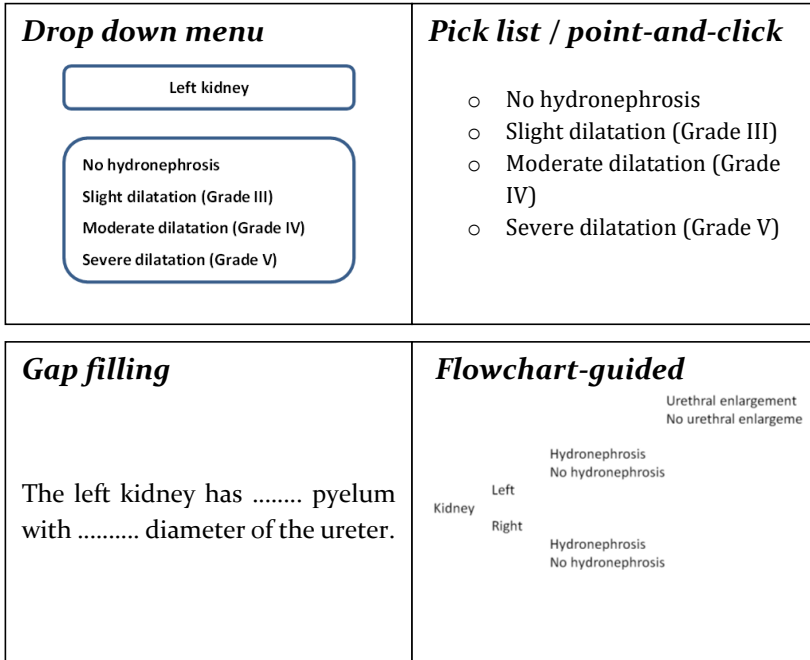


Figure 2. Structured reporting level 2: Structured content. Examples of structured content. In a drop down menu, the reporter chooses from several options in a different fashion than in a pick list or point-and-click system. Gap filling allows the reporter to fill in the blanks, whereas in a flowchart-guided report options are followed by a certain input made earlier in the reporting process.

Discussion

The recent literature seems to classify any and every change in generating radiological reports as structured reporting. The lack of a clear definition therefore has led to widespread confusion between the terms standardization and structured reporting. By distinguishing standardization, with level 1 and level 2 structured reporting as separate concepts, it becomes clearer that structured reporting is more than simply changing the radiological report. We argue that it is critical to distinguish these three concepts, because each tackles the problem of improving reporting in radiology at another level.

Structured reporting should by definition include an IT-based tool or system supporting the reporter when creating the actual report and can be supported by an IT-based tool

that orders the report into a certain lay-out (level 1), or can be constructed by an IT-based tool that inserts predefined medical content (level 2).

Although the final radiological report may be identical in terms of readability and clarity regardless which structured reporting method has been used, it is important to realize that the choice of a specific IT tool to create the report significantly influences future data mining possibilities. Reports that are generated with drop down menus (level 2) can be mined with minimal effort, because outcomes (options) are already stored as structured data. However, reports that are created with only level 1 structured reporting (e.g. hierarchical structure or reports with subheadings) may be more difficult to mine, because data elements are stored with less structure or as non-structured free text. Therefore, the choice for a specific type of structured reporting should also be determined by the intended data mining target and the data mining method.

Standardization, on the other hand, is not a tool that supports the reporting process itself, but is an agreement about the content of the report in order to enhance its uniformity when implemented. This enhances the idea that standardization needs to be implemented before structured reporting to benefit clinical practice most. In other words, the medical content should be clearly defined and streamlined first, before it can be incorporated into an IT-based system facilitating structured reporting. Moreover, also standardization facilitates data mining by enhancing the consistency of used vocabulary.

Furthermore, this two-tiered definition provides a clear distinction between the clinical and IT-based challenges that must be overcome to improve reporting in radiology. Standardization should be developed in clinical practice, whereas structured reporting is developed by or in collaboration with vendors of IT-based reporting tools.

Currently, it seems that developments in the field of structured reporting are driven more by intuition, rather than actual scientific evidence. To our mind, reliable, evidence-based recommendations for implementing structured reporting can only be obtained by distinguishing between – and separately evaluating – standardization and structured reporting. Only proper differentiation between these concepts improves dedicated research, enables pooling and analysis of published data, and allows for proper implementation.

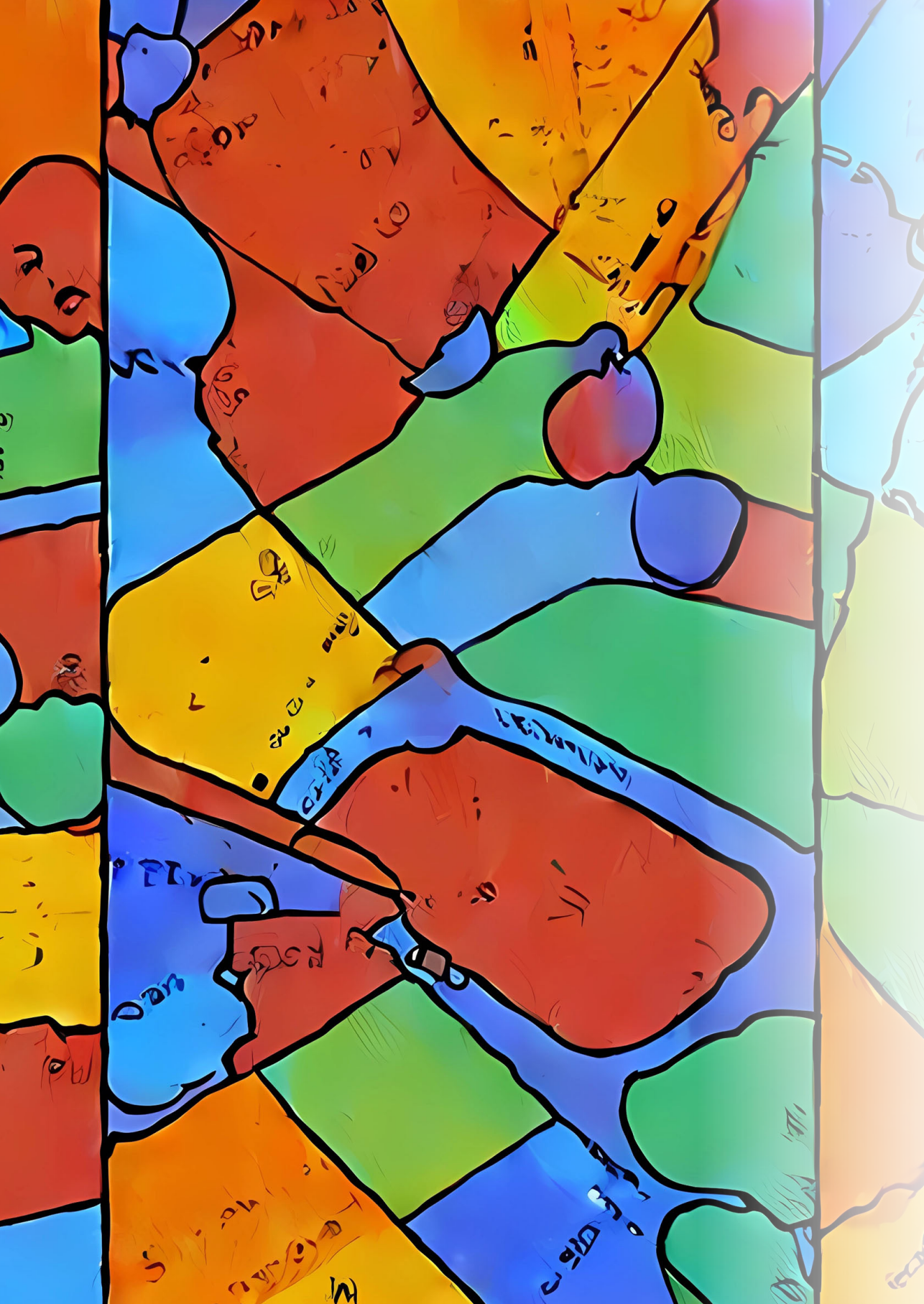
Conclusion

When incorporating structured reporting in clinical practice, it is important to consider its different forms, specific targets as well as its specific demands. In combination with proper standardization, the value of the radiological report can increase and data mining can be facilitated. Research and implementation should focus on the separate effects of standardized reporting and structured reporting, as both have its own value and impact in the process of reporting.

References

1. Reiner BI. The challenges, opportunities, and imperative of structured reporting in medical imaging. *J Digit Imaging*. 2009;22(6):562-8. doi: 10.1007/s10278-009-9239-z.
2. European Society of Radiology (ESR). ESR paper on structured reporting in radiology. *Insights Imaging*. 2018;9(1):1-7. doi: 10.1007/s13244-017-0588-8.
3. Powell DK, Silberzweig JE. State of Structured Reporting in Radiology, a Survey. *Acad Radiol*. 2015;22:226-33. doi: 10.1016/j.acra.2014.08.014.
4. Weiss DL, Bolos PR. Reporting and dictation. In: Branstetter IV BF, editor. *Practical imaging informatics: foundations and applications for PACS professionals*. New York: Springer; 2009. p. 147-162.
5. Hickey PM. Standardization of roentgen-ray reports. *AJR Am J Roentgenol*. 1922;9:442-445.
6. D'Orsi CJ, Sickles EA, Mendelson EB, Morris EA. *ACR BI-RADS® Atlas: Breast Imaging Reporting and Data System*. Reston (VA): American College of Radiology; 2013.
7. Turkbey B, Choyke PL. PIRADS 2.0: what is new? *Diagn Interv Radiol*. 2015;21(5):382-384.
8. Radiological Society of North America. *RadLex 3.12* [Internet]. Oak Brook (IL): Radiological Society of North America; 2016 [cited 2 Dec 2018] Available from <http://radlex.org>
9. Barbosa F, Maciel LM, Vieira EM, Azevedo Marques PM, Elias J, Muglia VF. Radiological reports: a comparison between the transmission efficiency of information in free text and in structured reports. *Clinics (Sao Paulo)*. 2010;65(1):15-21. doi: 10.1590/S1807-59322010000100004.
10. Liu D, Zucherman M, Tulloss Jr. WB. Six characteristics of effective structured reporting and the inevitable integration with speech recognition. *J Digit Imaging*. 2006;19:98-104. doi: 10.1007/s10278-005-8734-0.
11. Kahn CE Jr, Langlotz CP, Burnside ES, Carrino JA, Channin DS, Hovsepian DM, et al. Toward best practices in radiology reporting. *Radiology*. 2009;252(3):852-6. doi: 10.1148/radiol.2523081992.
12. Sistrom CL, Langlotz CP. A framework for improving radiology reporting. *J Am Coll Radiol*. 2005;2:159e67. doi: 10.1016/j.jacr.2004.06.015.

- 2
13. Weiss DL, Kim W, Branstetter IV BF, Prevedello LM. Radiology reporting: a closed-loop cycle from order entry to results communication. *J Am Coll Radiol*. 2014;11(12):1226-37. doi: 10.1016/j.jacr.2014.09.009.
 14. European Society of Radiology (ESR). Good practice for radiological reporting. Guidelines from the European Society of Radiology (ESR). *Insights Imaging*. 2011;2(2):93-96. doi:10.1007/s13244-011-0066-7.
 15. Krupinski EA, Hall ET, Jaw S, Reiner B, Siegel E. Influence of radiology report format on reading time and comprehension. *J Digit Imaging*. 2012;25(1):63-9. doi: 10.1007/s10278-011-9424-8.
 16. Karim S, Fegeler C, Boeckler D, Schwartz LH, Kauczor HU, von Tengg-Kobligk H. Development, implementation, and evaluation of a structured reporting web tool for abdominal aortic aneurysms. *JMIR Res Protoc*. 2013;2(2):e30. doi: 10.2196/resprot.2417.
 17. Hawkins CM, Hall S, Zhang B, Towbin AJ. Creation and implementation of department-wide structured reports: an analysis of the impact on error rate in radiology reports. *J Digit Imaging*. 2014 Oct;27(5):581-7. doi: 10.1007/s10278-014-9699-7.
 18. Kahn CE Jr, Wang K, Bell DS. Structured entry of radiology reports using World Wide Web technology. *Radiographics*. 1996 May;16(3):683-91. doi: 10.1148/radiographics.16.3.8897632.



Chapter 3:

Structured reporting in radiology: a systematic review to explore its potential

J. Martijn Nobel, Koos van Geel, Simon G.F. Robben

European Radiology (2021)

Abstract

Objectives: Structured reporting (SR) in radiology reporting is suggested to be a promising tool in clinical practice. In order to implement such an emerging innovation, it is necessary to verify that radiology reporting can benefit from SR. Therefore, the purpose of this systematic review is to explore the level of evidence of structured reporting in radiology. Additionally, this review provides an overview on the current status of SR in radiology.

Methods: A narrative systematic review was conducted, searching PubMed, Embase and the Cochrane Library using the syntax 'radiol*' AND 'structur*' AND 'report*'. Structured reporting was divided in SR level 1, structured lay-out (use of templates and checklists), and SR level 2, structured content (a drop down menu, point-and-click or clickable decision trees). Two reviewers screened the search results and included all quantitative experimental studies that discussed SR in radiology. A thematic analysis was performed to appraise the evidence level.

Results: The search resulted in 63 relevant full text articles out of a total of 8561 articles. Thematic analysis resulted in 44 SR level 1 and 19 level 2 reports. Only one paper was scored as highest level of evidence, which concerned a double cohort study with randomized trial design.

Conclusion: The level of evidence for implementing SR in radiology is still low and outcomes should be interpreted with caution.

Introduction

The area of radiology is an ever-innovating field with new applications, such as speech recognition systems and the introduction of Picture Archiving and Communication System (PACS), leading to digitalization and new possibilities in radiology reporting [1, 2]. The recent introduction of different types of structured reporting (SR) further accelerates initiatives in the field of reporting and many radiology departments use some sort of SR already [3]. The magnitude of this trend and its promotion by large radiological societies, such as the Radiological Society of North America (RSNA) and the European Society of Radiology (ESR), suggests that this way of reporting is

promising and that implementation of SR in clinical practice should be seriously considered [4, 5]. Overall, SR has been thought to be the key to improve clinical and radiological workflow.

The main goal of implementing SR seems to be enhancing the content of the radiological report as well as the reporting process itself. Due to increasing imaging possibilities, larger data sets and the availability of more specific treatments, details become ever more important. The radiological report should arrange this huge amount of information into a readable (legible) text containing the most accurate and specific information that is needed to make accurate decisions to treat the patient best. This renders the radiological reporting process more complicated and time consuming.

To accommodate this increasing demand of information, several tools have been proposed to improve the quality of the radiological report. Standardization tools (RECIST (Response Evaluation Criteria in Solid Tumors), Fleischner glossary, the RADS (Reporting And Data System) collection) [6, 7, 8], are created to be more accurate on describing pathology and its extension or evolution, to ensure that the content of the report is accurate. On the other hand, reporting tools, such as structured reporting and reporting guidelines, are constructed in order to enhance the reporting process; this concept is in literature generally referred to as “structured reporting”.

However, before implementation of SR, it is necessary to provide evidence to justify its introduction and implementation in the clinical workflow with a systematic review. As there is a plethora of definitions and interpretations of SR present in literature, a clear definition had to be determined for this review. The definition “*structured reporting is an IT-based method to import and arrange the medical content into the radiological report*”, as coined by Nobel et al. [9], was used. The main purpose of this systematic review is to explore the level of evidence of structured reporting. Additionally, this review provides an overview on the current status of SR in radiology.

Materials and methods

A systematic search was conducted according to the Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) criteria [10] and results were further categorized using a thematic analysis approach [11]. Results were analysed and

3

interpreted consistently with a textual narrative synthesis to visualize the similarities and differences among various methodologies in study design [12]. The next step was to determine the level of evidence of the studies. Because of the heterogeneity in study design, the simplified grading system (level A/B/C) according to Siwek et al. [13] was used to determine the strength of evidence on which outcomes were based. Randomized controlled trials are considered level A. Level B studies consist of all other evidence except for expert opinions or commentaries, which is level C. The groups were ordered on publication year followed by an alphabetical order. In case of discrepancy, consensus was reached between two authors (JMN and KG).

Literature review protocol

A literature search was conducted by searching PubMed, Embase and the Cochrane Library up to 10 August 2020. To include relevant papers, a wide search strategy was applied using the combination of the synonyms of 'radiology', 'structure' and 'reporting' (radiol* AND structur* AND report*).

Eligibility and study selection

All quantitative experimental studies that discussed SR in radiology have been included. After removing duplicates, title and abstract were independently screened on relevance by two authors. The following articles were excluded: articles that did not discuss structured reporting in radiology, comments or expert opinions (Level C [13]), articles not in English, German or Dutch, or those without full text availability. Bibliographies of included studies were searched in order to find additional relevant papers.

Definition of Structured Reporting (SR)

The definition "*structured reporting is an IT-based method to import and arrange the medical content into the radiological report*" [9], was used to frame the field of interest. This definition acknowledges a difference between SR and standardized reporting. Standardized reporting refers to the increase of uniformity of the report content with standardization tools (e.g. RECIST, Fleischner glossary, the RADS collection [6, 7, 8]). SR refers to the use of specific tools (structured reporting or reporting guidelines) that can be used to properly build, structure or fill the radiological report itself. This

differentiation is necessary to be able to only include the right studies which change the reporting process and not studies that merely change, for instance, the vocabulary used. Additionally, SR is subdivided into structured lay-out (SR level 1) and structured content (SR level 2) [9]. In this stratification model, structured lay-out (SR level 1) is defined as being a template or blueprint format in which the reporter has to report or has to adjust to. Structured content (SR level 2) is a manner in which the content of the radiology report can be inserted and displayed into the report (Fig. 1). As such, structured lay-out (e.g. templates and checklists), and structured content (e.g. drop down menu, point-and-click or clickable decision trees) highlight the level of IT involvement when implementing SR. This subdivision is used to be able to categorize the types of SR found in the included studies.

Structured lay-out SR level 1	<p>Itemized</p> <p>Abdomen: Liver: Gallbladder: Kidneys: Aorta:</p> <p>Impression:</p>	<p>Itemized – checklist</p> <p>Abdomen: Normal aspect of the liver: no; mass in the liver Normal aspect of the gallbladder: yes / no Normal aspect of the kidneys: yes / no Normal size of the aorta: yes / no</p> <p>Impression: Mass in the liver</p>
	<p>Drop down menu</p> <div style="border: 1px solid black; padding: 5px; margin-bottom: 5px; text-align: center;">Liver</div> <div style="border: 1px solid black; padding: 5px; text-align: center;"> Cyst Mass Metastasis Unknown </div>	<p>Point-and-click / pick list</p> <div style="display: flex; align-items: flex-start;"> <div style="margin-right: 20px;">Liver</div> <ul style="list-style-type: none"> <input type="radio"/> Segment 1 <input type="radio"/> Segment 2 <input type="radio"/> Segment 3 <input type="radio"/> Segment 4a <input type="radio"/> Segment 4b <input type="radio"/> Segment 5 ✓ <input type="radio"/> Segment 6 <input type="radio"/> Segment 7 <input type="radio"/> Segment 8 </div>

Figure 1. Examples of different levels of structured reporting; these are examples of IT-based tools to insert specific textual items into the radiological report, for instance with the use of a drop down menu in which an option can be chosen out of a particular list, or by using a point-and-click / pick list which in turn can open a new point-and-click/pick list option in order to build the report. SR level 1 = Structured lay-out: itemized, itemized-checklist; in these examples the obligated items or possible options are already stated in the template to ensure its presence. SR level 2 = Structured content: drop down menu, point-and-click/pick list.

Results

The literature search retrieved 4233, 6746 and 173 articles (total 11152) from PubMed, Embase and the Cochrane Library databases respectively. 2591 duplicates were removed. Title and abstract of 8561 articles were assessed by JMN and KG on, which resulted in 58 relevant articles. Full text was available for 56 articles. Bibliography search resulted in 7 additional studies, leading to a total of 63 studies that were included (Fig. 2, Table 1). No reviews were found. Due to the heterogeneity of included studies, it was neither possible to perform a meta-analysis nor to pool the results.

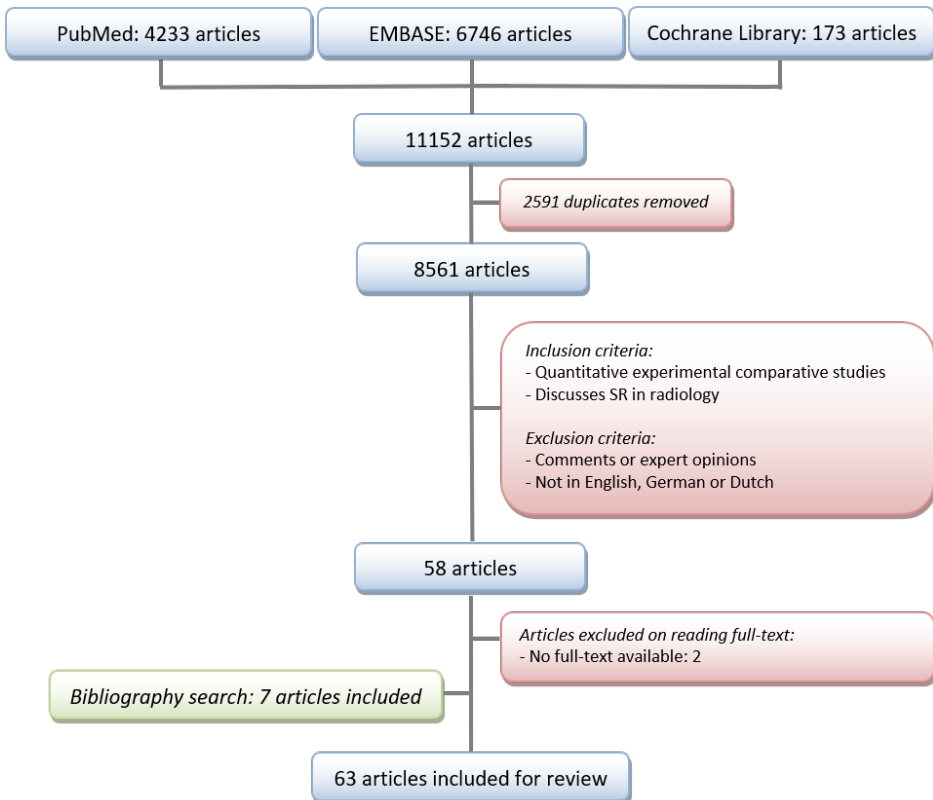


Figure 2. Search flow chart. SR = structured reporting

Level of evidence		Control	Intervention	Subspecialty/ field	Indication	Modality	Outcome(s)
Structured layout (SR level 1) – One Template							
	B	Free text	Structured itemized template with four parts and several key items	Abdomen	Pancreatic ductal adenocarcinoma	CT	Significant reduction of missing morphological and vascular features Improvement inter-reader agreement
Dimarco et al. (2020)¹⁴							
	B	Free text	Added 14 essential parameters	Abdomen	Rectal cancer staging	MRI	Significant report quality improvement Referring provider satisfaction improved
Gupta et al. (2020)¹⁵							
	B	Free text	Free form structured itemized templates	Abdomen	Various	CT	Less reporting errors potentially reducing The report word length did not differ
McFarland (2020)¹⁶							
	B	Free text	Additional template with key items for critical findings	Neurology	CNS metastasis	MRI	Automated insertion of context-dependent data and required elements is feasible Guideline adherence concerning critical findings improved
Olthof et al. (2020)¹⁷							
	B	Free text	Adding key features concerning inherited neuromuscular disorders	Musculoskeletal radiology	Lower limb inherited neuromuscular disorder	MRI	More clinically relevant disease management information
Alessandrino et al. (2019)¹⁸							
	B	Free text	Structured template with three options to score CNS metastasis after RT	Neurology	CNS metastasis	MRI	Decreasing non-specific description Improving discrete characterization
Benson et al. (2019)¹⁹							

Gore et al. (2019) ²⁰	B	Free text	Template with headings according to BT-RADS	Neurology	Brain tumor (BT-RADS)	MRI	Usage of non-specific language usage did not differ Perception improvement among radiologists and referring providers
Liu et al. (2019) ²¹	B	Free text	Structured itemized template with key features and standardized entries	Abdomen	Endometrial cancer	MRI	Increasing radiologists' work efficiency and gynaecologists' satisfaction
Wetterauer et al. (2019) ²²	B	Free text	Structured reports with PI-RADS key features	Abdomen	Prostate cancer (PI-RADS)	MRI	Urologists' surgical planning was facilitated by better assessing exact tumor location Improved satisfaction referring physician
Bink et al. (2018) ³³	B	Free text	Itemized template (17 tumor items)	Neurology	Brain tumor staging	MRI	Template ensured reliable detection of all relevant predefined items and reproducible documentation
Griffin et al. (2018) ⁴	B	Free text	Itemized template with TI-RADS and/or management integration	Head and Neck	Thyroid nodules (TI-RADS)	Ultrasound	Better feature description ACR TIRADS usage substantially improved management recommendations
Magnetta et al. (2018) ²⁵	B	Free text	Itemized template using PI-RADS	Abdomen	Prostate (PI-RADS)	MRI	Improved communication and clinical report impact with referring urologists
Olthof et al. (2018) ²⁶	B	Free text	Itemized RECIST template	Various	RECIST	CT	Combination of optimized workflow, subspecialization and SR led to significantly better report quality
Poulios et al. (2018) ²⁷	B	Free text	Itemized template	Abdomen	Hepatocellular carcinoma	CT	Assessment of transplant suitability improved using Milan criteria

Tersteeg et al. (2018) ³⁸	B	Free text	Itemized template with incorporated guidelines and key features	Abdomen	Rectal cancer staging	MRI	More complete report
Flusberg et al. (2017) ³⁹	B	Free text	Itemized template incorporating including LI-RADS	Abdomen	Hepatocellular carcinoma (LI-RADS)	MRI / CT	More comprehensive and consistent reporting
Franconeri et al. (2017) ³⁰	B	Free text	Disease specific itemized template	Abdomen	Uterine fibroid	MRI	Fewer key features were missed More helpful for treatment planning & understanding Improved reimbursement
Pysarenko et al. (2017) ³¹	B	Free text	Template with 8 itemized key-elements	Abdomen	Various	Ultrasound	
Wildman-Tobriner et al. (2017) ³²	B	Free text	Itemized template	Abdomen	IBD	CT	Key feature reporting improved Minimal impact on accuracy SR reports were preferred by referring physicians
Wildman-Tobriner et al. (2017) ³³	B	Free text	Itemized template with 15 key elements	Abdomen	Pediatric Crohn's disease	MRI	Significantly increasing on key features mentioning Referring clinicians subjectively preferred SR
Dickerson et al. (2016) ³⁴	B	Free text	Itemized template with 12 key features	Brain	MS	MRI	Increased rate relevant findings Standardized reports are preferred by neurologists
Brook et al. (2015) ³⁵	B	Free text	Itemized template with 12 key features	Abdomen	Pancreatic cancer	CT	Superior evaluation Facilitated surgical planning Increased surgeons' confidence concerning tumor resectability
Sahni et al. (2015) ³⁶	B	Free text	Template with 14 itemized quality measures	Abdomen	Rectal cancer staging	MRI	Report quality improved, 30% of reports remained unsatisfactory

Silveira et al. (2015)³⁷	B	Free text	Itemized template and computer-aided diagnosis	Abdomen	Prostate	MRI	Improving report quality Improving contrast enhancement kinetic curve
Lin et al. (2014)³⁸	B	Free text	Itemized checklist-based template	Neurology / Trauma	Cervical spine	CT	Significant decrease in missed non-fracture findings No change in missed fractures
Marcovici et al. (2014)³⁹	B	Free text	Prepopulated itemized checklist template	Thorax	Various	X-ray	Templates are more complete and more effective
Powell et al. (2014)⁴⁰	B	Free text	Itemized checklist-based template	Neurology / Trauma	Maxillofacial	CT	No improvement on report accuracy of radiology residents Focused training, checklist flexibility, and an adjustment period are important Only mandatory checklists were readily adopted by residents
Fraser et al. (2013)⁴¹	B	Free text	Itemized template with different options (paper)	Head and Neck	Cervical lymphadenopathy	Ultrasound	Increased report streamline
Structured layout (SR level 1) – Multiple templates							
Chung et al. (2020)⁴²	B	Free text	Seven different cross-divisional standardized structured reports	Thorax	Various	X-ray	Improvement of economic gains and projected radiologist time
Hanna et al. (2016)⁴³	B	Free text	Seven different itemized templates (4 CTs, 2 X-rays, 1 ultrasound)	Emergency	Various	Various	Decrease of dictation time Decrease of total word length in some cases Mixed impact on total reporting time

Hawkins et al. (2014) ⁴⁴	B	Free text	228 different prepopulated templates which may consist a pick list, fill-in-field and/or prose dictation	Various	Various	Various	Carefully constructed structured reports can help reducing errors
Larson et al. (2013) ⁴⁵	B	Free text	228 different prepopulated templates which may consist a pick list, fill-in-field and/or prose dictation	Various	Various	Various	High implementation adaptation rate
Hawkins et al. (2012) ⁴⁶	B	Free text	Different prepopulated templates	Various	Various	Various	Prepopulated reports alone do not affect error rate or dictation time of radiology reports
Schwartz et al. (2011) ⁴⁷	B	Free text	Different itemized templates	Various	Various	CT	Better content and greater clarity for radiologists and referring clinicians
Liu et al. (2003) ⁴⁸	B	Free text	Different menu-based templates	Various	Various	Various	Faster report turn-around time Less transcription errors and lower transcription costs
Structured layout (SR level 1) – Hypothetical research							
Dabrowiecki et al. (2020) ⁴⁹	B	Free text	One negative chest X-ray report compared with one out of four templates	Thorax	Chest	X-ray	Template use resulted in better comprehension by the public Unnecessary follow-up was less likely
Camilo et al. (2019) ⁵⁰	B	Free text	Four different templates (one free text, two ultrasound and one CT report)	Abdomen	Various	Ultrasound CT	Structured report with final conclusion / comment is preferred by attending and requesting physicians
Heye et al. (2018) ⁵¹	B	Free text	Three different layouts (structured itemized text, tables, images)	Thorax	Chest	CT	The customer favors structured reporting
Lather et al. (2017) ⁵²	B	Free text	Structured itemized template	Thorax	Chest	CT	SR is superior

Travis et al. (2014) ⁵³	B	Free text	Three different layouts with measurement section	Thorax / Abdomen	Various oncological	CT	A separate lesion measurement section is preferred over random mentioning
Krupinski et al. (2011) ⁵⁴	B	Free text	Itemized and hierarchical template	Abdomen	Renal abnormalities	CT	A "one-size-fits-all" radiology report format does not exist
Grieve et al. (2008) ⁵⁵	B	Free text	Four different templates	Abdomen	Negative examination	Ultrasound	Detailed reports and a radiologists' opinion is preferred by general practitioners
Sistrom et al. (2005) ⁵⁶	B	Free text	Itemized structured templates	Abdomen	Renal calcifications	CT	Equally efficient and accurate for transmitting content
Naik et al. (2001) ⁵⁷	B	Free text	Three itemized with difference in completeness	Abdomen	Various	Ultrasound	Improved facilitation of complete documentation
Structured content (SR level 2)							
Johnson et al. (2010) ^{58a}	A	Free text	Point-and-click system used to build a sentence in the structured report	Neurology	Possible stroke	MRI	No improvement in report clarity by attending physicians
Johnson et al. (2009) ^{59a}	A	Free text	Point-and-click system used to build a sentence in the structured report	Neurology	Possible stroke	MRI	Report accuracy and completeness did not improve
Aase et al. (2020) ⁶⁰	B	Free text	Template checklist with six pick list options concerning incidental pulmonary nodule description	Thorax	Pulmonary nodule	CT	Increased documentation compliance Better follow-up process Low utilization rates

Alper et al. (2020)⁶¹	B	Free text	Template with pick list options with preferred terms for abdominal organs normal finding mentioning	Abdomen	Various	CT / MRI	Better use of preferred / acceptable phrases Decreased use of equivocal terms
Kim et al. (2020)⁶²	B	Free text	Template-based structured reports with point-and-click menus including standard elements used in a densitometry report	Nuclear radiology	Osteoporosis	DXA	Shorter reporting times Increased report quality
Tuncyurek et al. (2019)⁶³	B	Free text	Template with pick list options to describe 12 key features of pelvic MRI for perianal fistulizing disease	Abdomen	Perianal fistulizing disease	MRI	Fewer key features were missed More complete, clear and helpful for treatment planning
Armbruster et al. (2018)⁶⁴	B	Free text	Clickable decision trees that function as a checklist and to use for building automatically semantic sentences	Head and Neck	Petrous bone	MRI	Increases completeness and quality Satisfaction of referring physicians improved
Sabel et al. (2018)⁶⁵	B	Free text	Clickable decision trees on several items with several subitems concerning vascular status	Vascular	Lower extremities	CTA	Superior clarity, completeness, clinical relevance, and usefulness rated by referring clinicians
Schoeppe et al. (2018)⁶⁶	B	Free text	Clickable decision trees in which outcomes were used to create semantic sentences and were displayed in the report	Abdomen	Swallowing disorders	Swallowing studies	Increases detailed information and facilitation of information extraction Better assisting clinical decision-making
Schöppe et al. (2018)⁶⁷	B	Free text	Clickable decision trees for specific items concerning (degenerative) osteoarthritis of the glenohumeral joint used to create semantic sentences used in the report	Musculoskeletal radiology	Shoulder	X-ray	May be a useful tool in clinical decision-making
Shaish et al. (2018)⁶⁸	B	Layout template	Drop down menus which were used as template to describe individual lesion characteristics concerning PI-RADS	Abdomen	Prostate	MRI	PI-RADS adherence improved May increase diagnostic performance

Gassenmaier et al. (2017)⁶⁹	B	Free text	Template with findings and impression section with clickable decision trees with several levels	Musculoskeletal radiology	Shoulder	MRI	Improved readability Improved linguistic quality
Norenberg et al. (2017)⁷⁰	B	Free text	Clickable decision trees used to describe 13 key features	Abdomen	Rectal cancer	MRI	Facilitates surgical planning Higher satisfaction level of referring surgeons about report correctness and clinical decision making
Sabel et al. (2017)⁷¹	B	Free text	Clickable decision trees containing observations with standardized subheadings in a consistent order	Thorax	Pulmonary embolism	CTA	Superior in clarity, better content and clinical utility
Walter et al. (2015)⁷²	B	Free text	Pick list about coronary calcifications added to a structured report with normal and abnormal default standard terminology which auto-populates the report	Cardio	Coronary calcifications	CT	Improved accuracy of coronary calcification mentions
Schweitzer et al. (2014)⁷³	B	Free text	Template with 108 obligated items with drop down menus and free text option. The report contains highlighted parts when stated as abnormal	Forensics	Whole body	CT	Can act as guideline
Karim et al. (2013)⁷⁴	B	Free text	Different IT-based options were used and included standardized point-and-click menus, including anatomy, measures and additional diagnostic findings listed by organ and dedicated pathology in three different sections with a free text option for personal judgment	Vascular	Abdominal aortic aneurysm	CTA	Decrease in average reporting time Ease of use may lead to more accurate decision support.
Barbosa et al. (2010)⁷⁵	B	Free text	Pick list reporting system on 8 descriptive items necessary for thyroid nodule characterisation	Head and Neck	Thyroid	Ultrasound	Information transmission improved for radiologists and referring clinicians
Hasegawa et al. (2010)⁷⁶	B	Free text	Pick list items and particular modifiers for different categories can be entered in templates that link those together	Thorax	Chest	X-ray	Report production time decreased

Table 1. Study characteristics and overview of articles with level A and B evidence which studied structured reporting in radiology. Presented is the level of evidence, control group, intervention, subspecialty/field, indication, modality and outcome(s). ^a Identical study population or cohort.

SR = Structured Reporting; SR level 1 = Structured lay-out; SR level 2 = Structured content; CNS = Central Nervous System; BT-RADS = Brain Tumor-Reporting And Data System; PI-RADS = Prostate Imaging-Reporting And Data System; TI-RADS = Thyroid Imaging-Reporting And Data System; RECIST = Response Evaluation Criteria in Solid Tumours; LI-RADS = Liver Imaging-Reporting And Data System; RT = Radiotherapy; IBD = Irritable Bowel Disease; MS = Multiple Sclerosis

Thematic data analysis

After inclusion, the 63 studies were grouped into structured lay-out (SR level 1) and structured content (SR level 2) groups (Fig. 3).

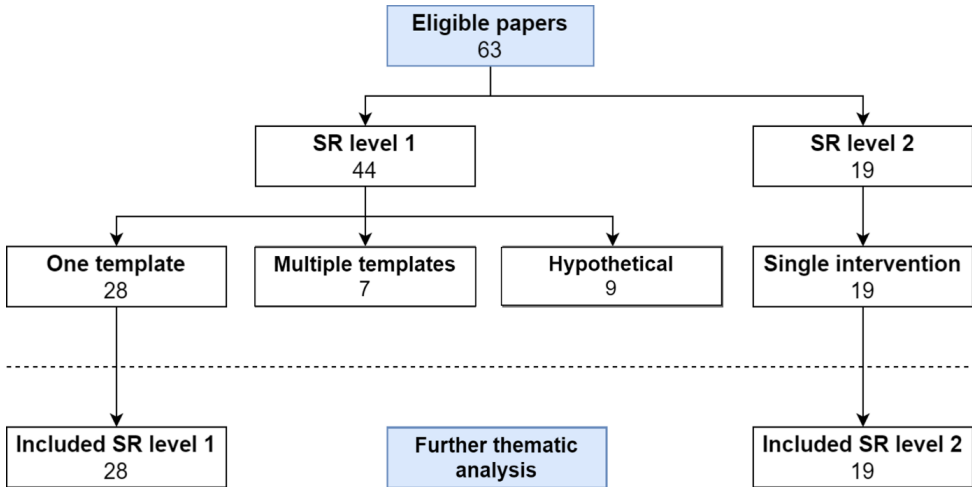


Figure 3. Characteristics of included studies based on SR level. SR level 1 = Structured lay-out; SR level 2 = Structured content

Control group, intervention, subspecialty/field, indication, modality and outcome of each study was assigned. Because of heterogeneity in the structured lay-out group (SR level 1), this group of 44 studies was subdivided into three subcategories: 1) one template ($n=28$), 2) multiple templates ($n=7$) and 3) hypothetical research ($n=9$) (Table 1, Fig. 3 and Fig. 4).

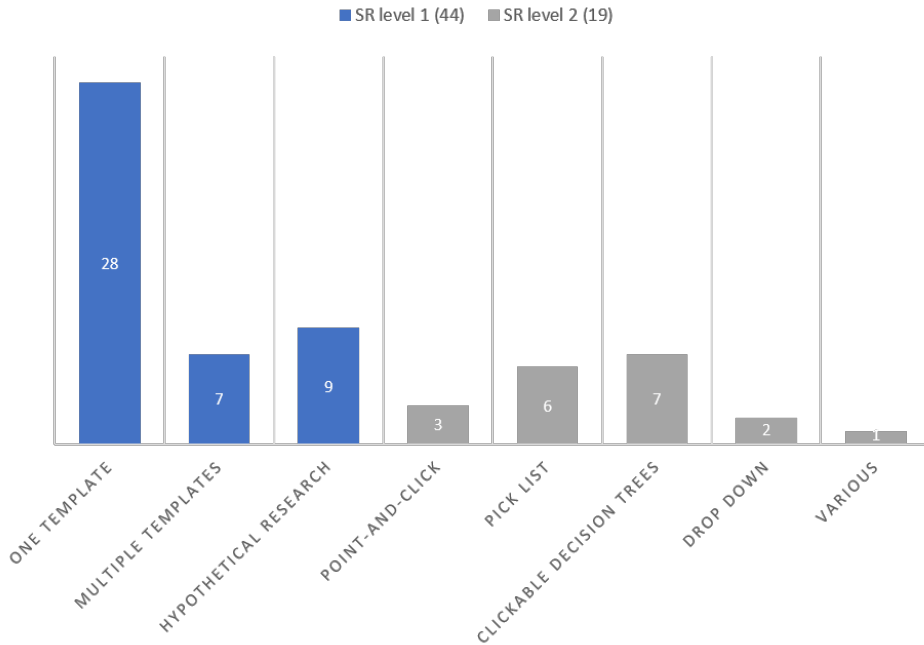


Figure 4. Intervention based on SR level. SR level 1 = Structured lay-out; SR level 2 = Structured content

The first subcategory “*one template*” consists of studies that implement and compare only one template with a free text report comparison. An example can be an itemized template to report a specific clinical question, such as a Magnetic Resonance Imaging (MRI) for brain tumor staging. The second subcategory “*multiple templates*” implemented several templates at once in their study before the comparison with free text reports was made. An example can be the implementation of several different templates for different clinical questions, such as implementing templates for Computed Tomography (CT), ultrasound and X-ray concerning kidney stones, appendicitis and heart failure. The third subcategory “*hypothetical research*” concerned studies that did not actually implement SR in clinical workflow, but assessed clinical or referring preferences on how to present the radiological information in the radiological report.

All 19 structured content (SR level 2) studies were interventional studies using an IT-based method to create the radiological report in the subcategories point-and-click

system, pick list, clickable decision trees, drop down and various (Table 1, Fig. 3 and Fig. 4).

As it is only possible, in an evidence-based manner, to accurately compare one structured reporting tool in one clinical interventional setting at once, only the studies implementing one template from the structured lay-out group and non-hypothetical studies have used for further analysis. When not taking into account the hypothetical studies, nor the studies of the multiple template category, 28 studies remain on the structured lay-out level (SR level 1). All 19 structured content (SR level 2) studies were interventional studies using one IT-based method to create the radiological report and were all suitable for further analysis (Table 1, Fig. 3 and Fig. 4). The remaining subcategories (one template SR level 1 and all SR level 2 studies) resulted in 47 studies (Fig. 3).

Further analysis of these 47 studies resulted in additional characteristics about subspecialty field and used modalities (Fig. 5a and Fig. 5b). Overall, CT and MRI modalities are mostly used on the subspecialties abdomen and neurology.

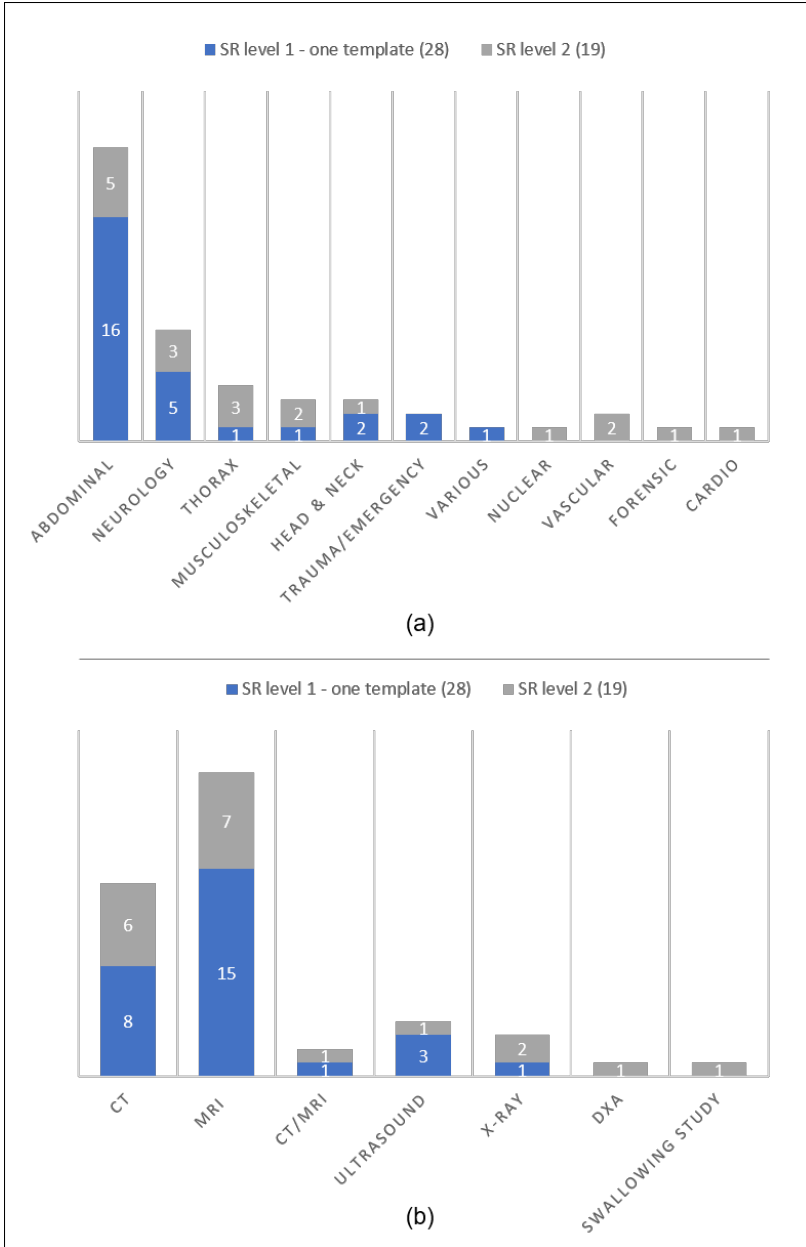


Figure 5. a) Subspecialty based on SR level and b) modality used based on SR level. All included single intervention studies according to the field of specialty and modality used. SR level 1 = Structured lay-out; SR level 2 = Structured content; DXA = Dual-energy X-ray absorptiometry (DXA)

Level of evidence

Two papers (one single study) were scored as level A in the structured content group. All other studies in the structured lay-out and structured content group were scored as level B evidence (Fig. 6).

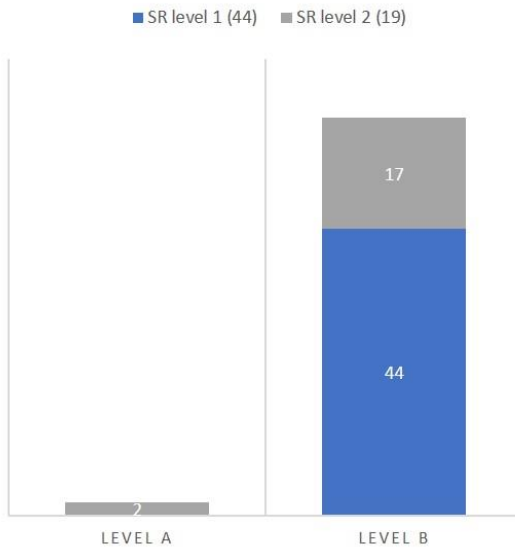


Figure 6. Level of evidence based on SR level. Level A = level A evidence according to Siwek et al. [13]; SR level 1 = Structured lay-out; SR level 2 = Structured content

Outcome

The value of outcomes of the studies on structured reporting depends heavily on the level of evidence of these studies. Therefore, the main focus of this study was to determine the level of evidence. However, to create an overview of research done on SR in radiology, main outcomes of included SR studies have been summarized in Table 1.

Discussion

The main goal of this narrative systematic literature review was to explore the level of evidence of all studies that try to enhance the radiological reporting process by using SR. This also resulted in an overview on the current status of SR in radiology and a

summary of its outcomes. To our knowledge, this is the first paper to provide a systematic review of SR in radiology.

Level of evidence

A double blinded, randomized controlled trial is considered as the highest level of original research (not including systematic reviews or meta-analysis). In our literature search, the only study that approximates this level was the double cohort study with randomized trial design conducted by Johnson et al. [58, 59] and was therefore scored as level A evidence. They compared a point-and-click reporting system (SR level 2) with free text reporting in brain MRI in stroke patients in two papers. This study states that only the way of reporting varied in order to exclude all other interfering factors, thereby only investigating the effect of the change in reporting method. The remaining 61 studies were considered level B evidence, showing an overall low level of evidence.

The hypothetical subcategory studies ($n=7$) are not implementational but only exploratory of nature. The multiple templates studies ($n=9$) are considered low level evidence, because it is virtually impossible to confidentially match outcomes to a particular way of reporting, when a) introducing several templates or reports simultaneously, b) using different levels of SR, for c) trying to answer different clinical questions.

However also the other subcategory studies (one template SR level 1 and all SR level 2 studies), except both level A studies, changed several factors during the implementation of SR, which again can result in some sort of confounding. For instance, many papers describe an expert meeting among radiologists and/or clinicians, or conducted a literature review in order to create a template or pick-list with adequate vocabulary, before implementing SR. This introduced an additional standardizing step next to the implementation of SR in the reporting routine. As a result, both the report content and the reporting manner differed, and outcomes of these studies reflect the effect of the combined interventions. The effects of any individual intervention, however, remain unclear.

Additionally, an expert meeting or literature review before implementing the new reporting manner will likely result in an increase in report quality or accuracy, because the reporter will be guided in stating the correct (newly stated) items necessary for

diagnosing when using SR, and thereby enhancing the report content. In this way, confirmation bias can occur, especially when report content quality or accuracy was the main goal of the study, and when outcomes were scored by the same experts that participated in the initial expert meeting.

The aforementioned shows that the study design of the included studies was hampered, resulting in low level of evidence studies. However, despite the fact that most studies are of low evidence, the total amount of published papers show the magnitude of the trend towards structured reporting in radiology.

One of the issues in chosen study design is probably based on the willingness to improve the radiological report as final clinical outcome, rather than searching for the true (single) vehicle that facilitates this.

Furthermore, a reason for the lack of high-level evidence papers can be the fact that proper implementation of SR might be highly case-specific. In radiology, multiple modalities as well as multiple clinical questions coexist and therefore it is possible that a SR tool or a specific SR level is not beneficial for all clinical settings or that it is depending on for instance difficulty level. A point-and-click or clickable decision tree method (SR level 2) may be better for a simple task with only few options, such as describing a thyroid nodule on an ultrasound examination. Likewise, a difficult, extensive clinical question which needs highly specific information or an extensive description, such as the description of a brain tumor on MRI, may suit a template or checklist (SR level 1) better than a point-and-click/pick list. In combination with several vendor dependent structuring methods on different SR levels, this makes it difficult to choose a specific topic to set up a well-designed study. Also the fact that there are no studies found that compare two different SR methods, but only comparing free text with some sort of SR, shows that research on SR in radiological reporting is still at an exploratory level.

Current standing and future perspectives

Looking at the levels of SR, in total, 28 studies were performed at the level of structured lay-out implementing one template and 19 on the structured content level implementing a more IT-based type of SR, which shows that both SR level 1 and 2 are used in clinical studies. It is interesting to see that both levels are being investigated,

because it is important to realize that in most cases it is easier, due to its lower IT-demand, to implement a template (SR level 1) in the reporting process than, for instance, implement a drop down menu-based report (SR level 2).

When looking at modality and subspecialty, most efforts are made with reports of CT and MRI examinations in the field of abdominal radiology and neuroradiology. An explanation might be the fact that the most important (staging) procedures use CT and MRI as a modality. Perhaps, the abdominal and neuroradiology fields are more suitable for using templates or it can be triggered by the fact that good classification systems or standardization systems already exist in these fields. If this is the case, this highlights the fact that SR is used for standardization by making sure that specific items or classification systems are described or used.

Table 1 shows that SR level 1 (templates) are mainly used to describe key features necessary to stage a particular disease or tumor with a predefined sentence with or without a particular standardization tool. Used standardization tools or classification systems can be found in table 1, and examples are for instance PI-RADS, LI-RADS and RECIST, but also key elements concerning Crohn's disease, rectal cancer staging, Multiple Sclerosis (MS), trauma or head and neck lymphadenopathy are used. Hence, also SR level 2 studies use key feature description or standardization tools (e.g. PI-RADS) to describe specific disease or tumors, such as stroke, pulmonary nodules, rectal cancer, thyroid nodules or prostatic cancer (Table 1). However, SR level 2 studies use an IT-based system that supports constructing (semantic) sentences, according to the chosen option from the drop down menu or point-and-click system, in which standardization is almost automatically linked to structured reporting.

When looking at the study outcomes in Table 1, the main goals, incentives, used SR method and outcomes of each study vary widely, and therefore, pooling of outcomes is difficult. Despite this heterogeneity this table of outcomes provides a panoramic overview of the present status of SR in radiology.

It shows that most of the included papers show an improvement in outcome when implementing SR. However, when looking at the evidence level, the only level A study [58, 59] did not improve the report clarity, accuracy and completeness of the report using their point-and-click method. This is an interesting finding and can show that this particular point-and-click system was not beneficial in radiological reporting in this

specific setting and concerning this specific outcome. However, the outcome of this study alone is insufficient to state that SR level 2 is not beneficial in radiology reporting, because outcomes seem to be highly case-specific. However, it is also hard to state that SR is beneficial in reporting in radiology when looking at the low level of evidence of all other included studies.

Overall, the level of evidence for SR is low and especially the link between structured reporting and standardization and its different effects on the radiological report is currently overlooked, but is of utmost importance. It seems that improving radiology reporting is more than just implementing SR and that standardization is necessary next to SR, and that both are highly entangled when implementing SR. This is likely caused by the fact that SR is based on a rather strict format in which several (mandatory) items or key features should be reported. Perhaps the question should be whether SR is not just a means to facilitate standardization, rather than that SR is improving the radiological report itself.

As such, high quality research is necessary to separately investigate the value of all individual factors that are involved in standardization and SR to determine the best type of SR for a specific clinical problem. Investigating the effect of standardization should be prioritized, because it may make sense that improving the content of the report, hence making a complete report with all items referring clinicians are asking for, will likely improve reporting quality. Then, the next question should be how this standardized information should be placed in the radiological report and how we can assure it is inserted correctly. For instance, this can be done with a simple template or checklist (SR level 1), or with a more sophisticated point-and-click system (SR level 2). Finally, it is important to know whether the efforts are beneficial for the patient (e.g. better staging), the referring clinician (e.g. reduced reading time), the reporter (e.g. faster reporting) or for all. Nevertheless, it is possible that this supposed reporting improvement is mainly caused by standardization rather than SR.

Limitations

First of all, it was difficult to find all relevant implementational studies published on the subject of SR due to ambiguous use of the terms 'standardized reporting' and 'structured reporting'. To be as complete as possible, as well as to answer the research question

best, a prior set definition for SR and its categorization system was used. In addition, a bibliography search was used to search for missed studies after conducting the main search. Because of heterogeneity of the included studies, it was hard to pool the data on a more specific level and therefore a thematic analysis was used. The outcome analysis performed in this paper was limited by the large heterogeneity of outcomes and study design. A more thorough analysis should be done to explore outcome measurements better and to see who (the referring clinician, radiologist or patient) will benefit from SR most, as well as which specific efforts resulted in this outcome.

Conclusion

Structured reporting is thought to have great potential to improve reporting in radiology. However, due to difficulties in study design there is a lack of high-quality research on this topic resulting in low overall evidence. Future research is needed to explore the individual effects of standardization and SR, as it is questionable whether SR is the solution for improving reporting in radiology or only a means in facilitating standardization.

References

1. Liu D, Zucherman M, Tulloss Jr. WB. Six characteristics of effective structured reporting and the inevitable integration with speech recognition. *J Digit Imaging.* 2006;19:98-104. doi: 10.1007/s10278-005-8734-0.
2. Reiner BI. The challenges, opportunities, and imperative of structured reporting in medical imaging. *J Digit Imaging.* 2009;22(6):562-8. doi: 10.1007/s10278-009-9239-z.
3. Powell DK, Silberzweig JE. State of Structured Reporting in Radiology, a Survey. *Acad Radiol.* 2015;22:226-33. doi: 10.1016/j.acra.2014.08.014.
4. Radiological Society of North America. RadReport template library [Internet]. Oak Brook (IL): Radiological Society of North America; 2020 [cited 15 Dec 2020] Available from <https://radreport.org>
5. European Society of Radiology (ESR). ESR paper on structured reporting in radiology. *Insights Imaging.* 2018;9(1):1-7. doi: 10.1007/s13244-017-0588-8.
6. Schwartz LH, Seymour L, Litière S, Ford R, Gwyther S, Mandrekar S, et al. RECIST 1.1 - Standardisation and disease-specific adaptations: Perspectives from the RECIST Working Group. *Eur J Cancer.* 2016;62:138-45. doi: 10.1016/j.ejca.2016.03.082.

7. Hansell DM, Bankier AA, MacMahon H, McLoud TC, Müller NL, Remy J. Fleischner Society: glossary of terms for thoracic imaging. *Radiology*. 2008;246(3):697-722. doi: 10.1148/radiol.2462070712.
8. An JY, Unsrdorfer KML, Weinreb JC. BI-RADS, C-RADS, CAD-RADS, LI-RADS, Lung-RADS, NI-RADS, O-RADS, PI-RADS, TI-RADS: Reporting and Data Systems. *Radiographics*. 2019;39(5):1435-1436. doi: 10.1148/rg.2019190087.
9. Nobel JM, Kok EM, Robben SGF. Redefining the structure of structured reporting in radiology. *Insights Imaging*. 2020;11(1):10. doi: 10.1186/s13244-019-0831-6.
10. Moher D, Liberati A, Tetzlaff J, Altman DG; PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *BMJ*. 2009;339:b2535. doi: 10.1136/bmj.b2535.
11. Mays N, Pope C, Popay J. Systematically reviewing qualitative and quantitative evidence to inform management and policy-making in the health field. *J Health Serv Res Policy*. 2005;10 Suppl 1:6-20. doi: 10.1258/1355819054308576.
12. Lucas PJ, Baird J, Arai L, Law C, Roberts HM. Worked examples of alternative methods for the synthesis of qualitative and quantitative research in systematic reviews. *BMC Med Res Methodol*. 2007;7:4. doi: 10.1186/1471-2288-7-4.
13. Siwek J, Gourlay ML, Slawson DC, Shaughnessy AF. How to write an evidence-based clinical review article. *Am Fam Physician*. 2002;65:251-8.
14. Dimarco M, Cannella R, Pellegrino S, Iadicola D, Tutino R, Allegra F, et al. Impact of structured report on the quality of preoperative CT staging of pancreatic ductal adenocarcinoma: assessment of intra- and inter-reader variability. *Abdom Radiol (NY)*. 2020;45(2):437-448. doi: 10.1007/s00261-019-02287-7.
15. Gupta NA, Mahajan S, Sumankumar A, Saklani A, Engineer R, Baheti AD. Impact of a standardized reporting format on the quality of MRI reports for rectal cancer staging. *Indian J Radiol Imaging*. 2020;30(1):7-12. doi: 10.4103/ijri.IJRI_308_19.
16. McFarland JA, Elkassem AMA, Casals L, Smith GD, Smith AD, Gunn AJ. Objective comparison of errors and report length between structured and freeform abdominopelvic computed tomography reports. *Abdom Radiol (NY)*. 2021;46(1):387-393. doi: 10.1007/s00261-020-02646-9.
17. Olthof AW, Leusveld ALM, de Groot JC, Callenbach PMC, van Ooijen PMA. Contextual Structured Reporting in Radiology: Implementation and Long-Term Evaluation in Improving the Communication of Critical Findings. *J Med Syst*. 2020;44(9):148. doi: 10.1007/s10916-020-01609-3.
18. Alessandrino F, Cristiano L, Cinnante CM, Tartaglione T, Gerevini S, Verdolotti T, et al. Value of structured reporting in neuromuscular disorders. *Radiol Med*. 2019;124(7):628-635. doi: 10.1007/s11547-019-01012-0.
19. Benson J, Burgstahler M, Zhang L, Rischall M. The value of structured radiology reports to categorize intracranial metastases following radiation therapy. *Neuroradiol J*. 2019;32(4):267-272. doi: 10.1177/1971400919845365.
20. Gore A, Hoch MJ, Shu H-KG, Olson JJ, Voloschin AD, Weinberg BD. Institutional Implementation of a Structured Reporting System: Our Experience with the Brain Tumor Reporting and Data System. *Acad Radiol*. 2019;26(7):974-980. doi: 10.1016/j.acra.2018.12.023.

21. Liu Y, Feng Z, Qin S, Yang J, Han C, Wang X. Structured reports of pelvic magnetic resonance imaging in primary endometrial cancer: Potential benefits for clinical decision-making. *PLoS One*. 2019;14(3):e0213928. doi: 10.1371/journal.pone.0213928.
22. Wetterauer C, Winkel DJ, Federer-Gsponer JR, Halla A, Subotic S, Deckart A, et al. Structured reporting of prostate magnetic resonance imaging has the potential to improve interdisciplinary communication. *PLoS One*. 2019;14(2):e0212444. doi: 10.1371/journal.pone.0212444.
23. Bink A, Benner J, Reinhardt J, De Vere-Tyndall A, Stieltjes B, Hainc N, et al. Structured Reporting in Neuroradiology: Intracranial Tumors. *Front Neurol*. 2018;9:32. doi: 10.3389/fneur.2018.00032.
24. Griffin AS, Mitsky J, Rawal U, Bronner AJ, Tessler FN, Hoang JK. Improved Quality of Thyroid Ultrasound Reports After Implementation of the ACR Thyroid Imaging Reporting and Data System Nodule Lexicon and Risk Stratification System. *J Am Coll Radiol*. 2018;15(5):743-748. doi: 10.1016/j.jacr.2018.01.024.
25. Magnetta MJ, Donovan AL, Jacobs BL, Davies BJ, Furlan A. Evidence-Based Reporting: A Method to Optimize Prostate MRI Communications With Referring Physicians. *AJR Am J Roentgenol*. 2018;210(1):108-112. doi: 10.2214/AJR.17.18260.
26. Olthof AW, Borstlap J, Roeloffzen WW, Callenbach PMC, van Ooijen PMA. Improvement of radiology reporting in a clinical cancer network: impact of an optimised multidisciplinary workflow. *Eur Radiol*. 2018;28(10):4274-4280. doi: 10.1007/s00330-018-5427-x.
27. Poullos PD, Tseng JJ, Melcher ML, Concepcion W, Loening AM, Rosenberg J, et al. Structured Reporting of Multiphasic CT for Hepatocellular Carcinoma: Effect on Staging and Suitability for Transplant. *AJR Am J Roentgenol*. 2018;210(4):766-774. doi: 10.2214/AJR.17.18725.
28. Tersteeg JJC, Gobardhan PD, Crolla RMPH, Kint PAM, Niers-Stobbe I, Boonman-de Winter LJM, et al. Improving the Quality of MRI Reports of Preoperative Patients With Rectal Cancer: Effect of National Guidelines and Structured Reporting. *AJR Am J Roentgenol*. 2018;210(6):1240-1244. doi: 10.2214/AJR.17.19054.
29. Flusberg M, Ganeles J, Ekinici T, Goldberg-Stein S, Paroder V, Kobi M, et al. Impact of a Structured Report Template on the Quality of CT and MRI Reports for Hepatocellular Carcinoma Diagnosis. *J Am Coll Radiol*. 2017;14(9):1206-1211. doi: 10.1016/j.jacr.2017.02.050.
30. Franconeri A, Fang J, Carney B, Justaniah A, Miller L, Hur HC, et al. Structured vs narrative reporting of pelvic MRI for fibroids: clarity and impact on treatment planning. *Eur Radiol*. 2018;28(7):3009-3017. doi: 10.1007/s00330-017-5161-9.
31. Pysarenko K, Recht M, Kim D. Structured Reporting: A Tool to Improve Reimbursement. *J Am Coll Radiol*. 2017;14(5):662-664. doi: 10.1016/j.jacr.2016.10.016.
32. Wildman-Tobriner B, Allen BC, Bashir MR, Camp M, Miller C, Fiorillo LE, et al. Structured reporting of CT enterography for inflammatory bowel disease: effect on key feature reporting, accuracy across training levels, and subjective assessment of disease by referring physicians. *Abdom Radiol (NY)*. 2017;42(9):2243-2250. doi: 10.1007/s00261-017-1136-1.

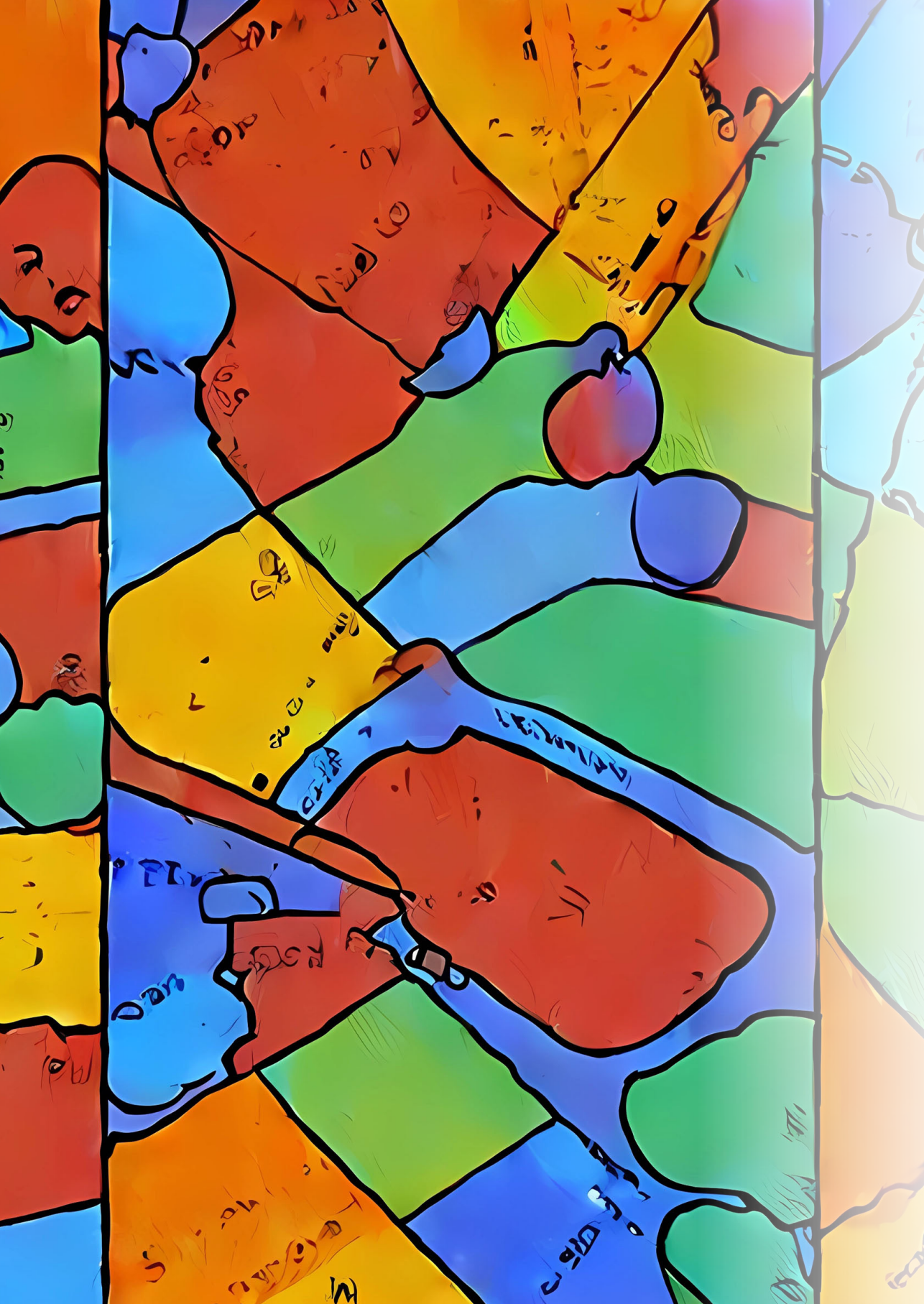
33. Wildman-Tobriner B, Allen BC, Davis JT, Miller CM, Schooler GR, McGreal NM, et al. Structured Reporting of Magnetic Resonance Enterography for Pediatric Crohn's Disease: Effect on Key Feature Reporting and Subjective Assessment of Disease by Referring Physicians. *Curr Probl Diagn Radiol.* 2017;46(2):110-114. doi: 10.1067/j.cpradiol.2016.12.001.
34. Dickerson E, Davenport MS, Syed F, Stuve O, Cohen JA, Rinker JR, et al. Effect of Template Reporting of Brain MRIs for Multiple Sclerosis on Report Thoroughness and Neurologist-Rated Quality: Results of a Prospective Quality Improvement Project. *J Am Coll Radiol.* 2017;14(3):371-379.e1. doi: 10.1016/j.jacr.2016.09.037..
35. Brook OR, Brook A, Vollmer CM, Kent TS, Sanchez N, Pedrosa I. Structured reporting of multiphase CT for pancreatic cancer: potential effect on staging and surgical planning. *Radiology.* 2015;274(2):464-72. doi: 10.1148/radiol.14140206.
36. Sahni VA, Silveira PC, Sainani NI, Khorasani R. Impact of a Structured Report Template on the Quality of MRI Reports for Rectal Cancer Staging. *AJR Am J Roentgenol.* 2015;205(3):584-8. doi: 10.2214/AJR.14.14053.
37. Silveira PC, Dunne R, Sainani NI, Lacson R, Silverman SG, Tempany CM, et al. Impact of an Information Technology-Enabled Initiative on the Quality of Prostate Multiparametric MRI Reports. *Acad Radiol.* 2015;22(7):827-33. doi: 10.1016/j.acra.2015.02.018.
38. Lin E, Powell DK, Kagetsu NJ. Efficacy of a checklist-style structured radiology reporting template in reducing resident misses on cervical spine computed tomography examinations. *J Digit Imaging.* 2014;27(5):588-93. doi: 10.1007/s10278-014-9703-2.
39. Marcovici PA, Taylor GA. Journal Club: Structured radiology reports are more complete and more effective than unstructured reports. *AJR Am J Roentgenol.* 2014;203(6):1265-71. doi: 10.2214/AJR.14.12636.
40. Powell DK, Lin E, Silberzweig JE, Kagetsu NJ. Introducing radiology report checklists among residents: adherence rates when suggesting versus requiring their use and early experience in improving accuracy. *Acad Radiol.* 2014;21(3):415-23. doi: 10.1016/j.acra.2013.12.004.
41. Fraser L, O'Neill K, Locke R, Attaie M, Irwin G, Kubba H, et al. Standardising reporting of cervical lymphadenopathy in paediatric neck ultrasound: a pilot study using an evidence-based reporting protocol. *Int J Pediatr Otorhinolaryngol.* 2013;77(8):1248-51. doi: 10.1016/j.ijporl.2013.04.026.
42. Chung CY, Makeeva V, Yan J, Prater AB, Duszak R Jr, Safdar NM, et al. Improving Billing Accuracy Through Enterprise-Wide Standardized Structured Reporting With Cross-Divisional Shared Templates. *J Am Coll Radiol.* 2020;17(1 Pt B):157-164. doi: 10.1016/j.jacr.2019.08.034.
43. Hanna TN, Shekhani H, Maddu K, Zhang C, Chen Z, Johnson J-O. Structured report compliance: effect on audio dictation time, report length, and total radiologist study time. *Emerg Radiol.* 2016;23(5):449-53. doi: 10.1007/s10140-016-1418-x.
44. Hawkins CM, Hall S, Zhang B, Towbin AJ. Creation and implementation of department-wide structured reports: an analysis of the impact on error rate in radiology reports. *J Digit Imaging.* 2014;27(5):581-7. doi: 10.1007/s10278-014-9699-7.

45. Larson DB, Towbin AJ, Pryor RM, Donnelly LF. Improving consistency in radiology reporting through the use of department-wide standardized structured reporting. *Radiology*. 2013;267(1):240-50. doi: 10.1148/radiol.12121502.
46. Hawkins CM, Hall S, Hardin J, Salisbury S, Towbin AJ. Prepopulated radiology report templates: a prospective analysis of error rate and turnaround time. *J Digit Imaging*. 2012;25(4):504-11. doi: 10.1007/s10278-012-9455-9.
47. Schwartz LH, Panicek DM, Berk AR, Li Y, Hricak H. Improving communication of diagnostic radiology findings through structured reporting. *Radiology*. 2011;260(1):174-81. doi: 10.1148/radiol.11010913.
48. Liu D, Berman GD, Gray RN. The use of structured radiology reporting at a community hospital: A 4-year case study of more than 200,000 reports. *Appl Radiol*. 2003;32:23-6.
49. Dabrowiecki A, Sadigh G, Duszak R. Chest Radiograph Reporting: Public Preferences and Perceptions. *J Am Coll Radiol*. 2020;17(10):1259-1268. doi: 10.1016/j.jacr.2020.04.003.
50. Camilo DMR, Tibana TK, Adórno IF, Santos RFT, Klaesener C, Gutierrez Junior W, et al. Radiology report format preferred by requesting physicians: prospective analysis in a population of physicians at a university hospital. *Radiol Bras*. 2019;52(2):97-103. doi: 10.1590/0100-3984.2018.0026.
51. Heye T, Gysin V, Boll DT, Merkle EM. JOURNAL CLUB: Structured Reporting: The Voice of the Customer in an Ongoing Debate About the Future of Radiology Reporting. *AJR Am J Roentgenol*. 2018;211(5):964-970. doi: 10.2214/AJR.18.19714.
52. Lather JD, Che Z, Saltzman B, Bieszczad J. Structured Reporting in the Academic Setting: What the Referring Clinician Wants. *J Am Coll Radiol*. 2018;15(5):772-775. doi: 10.1016/j.jacr.2017.12.031.
53. Travis AR, Sevenster M, Ganesh R, Peters JF, Chang PJ. Preferences for structured reporting of measurement data: an institutional survey of medical oncologists, oncology registrars, and radiologists. *Acad Radiol*. 2014;21(6):785-96. doi: 10.1016/j.acra.2014.02.008.
54. Krupinski EA, Hall ET, Jaw S, Reiner B, Siegel E. Influence of radiology report format on reading time and comprehension. *J Digit Imaging*. 2012;25(1):63-9. doi: 10.1007/s10278-011-9424-8.
55. Grieve FM, Plumb AA, Khan SH. Radiology reporting: a general practitioner's perspective. *Br J Radiol*. 2010;83(985):17-22. doi: 10.1259/bjr/16360063.
56. Sistrom CL, Honeyman-Buck J. Free text versus structured format: information transfer efficiency of radiology reports. *AJR Am J Roentgenol*. 2005;185(3):804-12. doi: 10.2214/ajr.185.3.01850804.
57. Naik SS, Hanbidge A, Wilson SR. Radiology reports: examining radiologist and clinician preferences regarding style and content. *AJR Am J Roentgenol*. 2001;176(3):591-8. doi: 10.2214/ajr.176.3.1760591.
58. Johnson AJ, Chen MYM, Zapadka ME, Lyders EM, Littenberg B. Radiology report clarity: a cohort study of structured reporting compared with conventional dictation. *J Am Coll Radiol*. 2010;7(7):501-6. doi: 10.1016/j.jacr.2010.02.008.
59. Johnson AJ, Chen MYM, Shannon Swan J, Applegate KE, Littenberg B. Cohort study of structured reporting compared with conventional dictation. *Radiology*. 2009;253(1):74-80. doi: 10.1148/radiol.2531090138.

60. Aase A, Fabbrini AE, White KM, Averill S, Gravely A, Melzer AC. Implementation of a Standardized Template for Reporting of Incidental Pulmonary Nodules: Feasibility, Acceptability, and Outcomes. *J Am Coll Radiol*. 2020;17(2):216-223. doi: 10.1016/j.jacr.2019.11.013.
61. Alper DP, Shinagare AB, Hashemi SR, Glazer DI, DiPiro PJ, Boland GW, et al. Effect of a Report Template-Enabled Quality Improvement Initiative on Use of Preferred Phrases for Communicating Normal Findings in Structured Abdominal CT and MRI Reports. *AJR Am J Roentgenol*. 2020;214(4):835-842. doi: 10.2214/AJR.19.21735.
62. Kim SH, Sobez LM, Spiro JE, Curta A, Ceelen F, Kampmann E, et al. Structured reporting has the potential to reduce reporting times of dual-energy x-ray absorptiometry exams. *BMC Musculoskelet Disord*. 2020;21(1):248. doi: 10.1186/s12891-020-03200-w.
63. Tuncyurek O, Garcés-Descovich A, Jaramillo-Cardoso A, Durán EE, Cataldo TE, Poylin VY, et al. Structured versus narrative reporting of pelvic MRI in perianal fistulizing disease: impact on clarity, completeness, and surgical planning. *Abdom Radiol (NY)*. 2019;44(3):811-820. doi: 10.1007/s00261-018-1858-8.
64. Armbruster M, Gassenmaier S, Haack M, Reiter M, Nörenberg D, Henzler T, et al. Structured reporting in petrous bone MRI examinations: impact on report completeness and quality. *Int J Comput Assist Radiol Surg*. 2018;13(12):1971-1980. doi: 10.1007/s11548-018-1828-1.
65. Sabel BO, Plum JL, Czihal M, Lottspeich C, Schönleben F, Gäbel G, et al. Structured Reporting of CT Angiography Runoff Examinations of the Lower Extremities. *Eur J Vasc Endovasc Surg*. 2018;55(5):679-687. doi: 10.1016/j.ejvs.2018.01.026
66. Schoeppe F, Sommer WH, Haack M, Havel M, Rheinwald M, Wechtenbruch J, et al. Structured reports of videofluoroscopic swallowing studies have the potential to improve overall report quality compared to free text reports. *Eur Radiol*. 2018;28(1):308-315. doi: 10.1007/s00330-017-4971-0.
67. Schöppe F, Sommer WH, Schmidutz F, Pförringer D, Armbruster M, Paprottka KJ, et al. Structured reporting of x-rays for atraumatic shoulder pain: advantages over free text? *BMC Med Imaging*. 2018;18(1):20. doi: 10.1186/s12880-018-0262-8.
68. Shaish H, Feltus W, Steinman J, Hecht E, Wenske S, Ahmed F. Impact of a Structured Reporting Template on Adherence to Prostate Imaging Reporting and Data System Version 2 and on the Diagnostic Performance of Prostate MRI for Clinically Significant Prostate Cancer. *J Am Coll Radiol*. 2018;15(5):749-754. doi: 10.1016/j.jacr.2018.01.034.
69. Gassenmaier S, Armbruster M, Haasters F, Helfen T, Henzler T, Alibek S, et al. Structured reporting of MRI of the shoulder - improvement of report quality? *Eur Radiol*. 2017;27(10):4110-4119. doi: 10.1007/s00330-017-4778-z.
70. Norenberg D, Sommer WH, Thasler W, D'Haese J, Rentsch M, Kolben T, et al. Structured Reporting of Rectal Magnetic Resonance Imaging in Suspected Primary Rectal Cancer: Potential Benefits for Surgical Planning and Interdisciplinary Communication. *Invest Radiol*. 2017;52(4):232-239. doi: 10.1097/RLI.0000000000000336.

71. Sabel BO, Plum JL, Kneidinger N, Leuschner G, Koletzko L, Raziorrouh B, et al. Structured reporting of CT examinations in acute pulmonary embolism. *J Cardiovasc Comput Tomogr.* 2017;11(3):188-195. doi: 10.1016/j.jcct.2017.02.008.
72. Walter WR, Goldberg-Stein S, Levsky JM, Cohen HW, Scheinfeld MH. A default normal chest CT structured reporting field for coronary calcifications does not cause excessive false-negative reporting. *J Am Coll Radiol.* 2015;12(8):783-7. doi: 10.1016/j.jacr.2015.03.011.
73. Schweitzer W, Bartsch C, Ruder TD, Thali MJ. Virtopsy approach: Structured reporting versus free reporting for PMCT findings. *J Forensic Radiol Imaging.* 2014;2:28-33. doi: 10.1016/j.jofri.2013.12.002
74. Karim S, Fegeler C, Boeckler D, Schwartz LH, Kauczor H-U, von Tengg-Kobligk H. Development, implementation, and evaluation of a structured reporting web tool for abdominal aortic aneurysms. *JMIR Res Protoc.* 2013;2(2):e30. doi: 10.2196/resprot.2417.
75. Barbosa F, Maciel LMZ, Vieira EM, Azevedo Marques PM de, Elias J, Muglia VF. Radiological reports: a comparison between the transmission efficiency of information in free text and in structured reports. *Clinics (Sao Paulo).* 2010;65(1):15-21. doi: 10.1590/S1807-59322010000100004.
76. Hasegawa Y, Matsumura Y, Mihara N, Kawakami Y, Sasai K, Takeda H, et al. Development of a system that generates structured reports for chest x-ray radiography. *Methods Inf Med.* 2010;49(4):360-70. doi: 10.3414/ME09-01-0014.

PART TWO:
NATURAL LANGUAGE PROCESSING



Chapter 4:

Natural Language Processing in Dutch free text radiology reports: challenges in a small language area staging pulmonary oncology

J. Martijn Nobel, Sander Puts, Frans C.H. Bakers,

Simon G.F. Robben, André L.A.J. Dekker

Journal of Digital Imaging (2020)

Abstract

Reports are the standard way of communication between the radiologist and the referring clinician. Efforts are made to improve this communication by, for instance, introducing standardization and structured reporting. Natural Language Processing (NLP) is another promising tool which can improve and enhance the radiological report by processing free text. NLP as such adds structure to the report and exposes the information, which in turn can be used for further analysis.

This paper describes pre-processing and processing steps and highlights important challenges to overcome in order to successfully implement a free text mining algorithm using NLP tools and machine learning in a small language area, like Dutch.

A rule-based algorithm was constructed to classify T-stage of pulmonary oncology from the original free text radiological report, based on the items tumor size, presence and involvement according to the 8th TNM classification system. PyContextNLP, spaCy and regular expressions were used as tools to extract the correct information and process the free text.

Overall accuracy of the algorithm for evaluating T-stage was 0,83 in the training set and 0,87 in the validation set, which shows that the approach in this pilot study is promising. Future research with larger datasets and external validation is needed to be able to introduce more machine learning approaches and perhaps to reduce required input efforts of domain-specific knowledge. However, a hybrid NLP approach will probably achieve the best results.

Introduction

One of the most challenging tasks in healthcare informatics nowadays is how to improve accessibility to medical information. Especially in radiology, in which a large amount of imaging and textual data is captured. Combining all kinds of medical information can improve current medical data flow and can ensure better healthcare [1]. A good example of a complex process of combining data is tumor staging, for instance in pulmonary oncology. A specific rule-based tumor classification system is used for

proper staging of pulmonary oncology, as stated in the 8th TNM Classification of Malignant Tumors (TNM) [2][3].

In radiology, the report is still considered the golden standard in communicating findings and is, despite several structuring efforts [4], usually still stored as free text. One of the challenges in radiology is how to (re-)use free text unstructured data of the radiological report for data mining purposes in, for instance, pulmonary tumor staging. Natural Language Processing (NLP) is a promising method for extracting information from free text, and has been used in several studies to extract data from radiological reports [5]. However, most use English as a language and specific medical NLP software, such as medical extraction systems (e.g. cTAKES) [6], are not available in Dutch [5][7]. In English, a rule-based pulmonary oncology TNM classification algorithm has already been built and trained on pathology reports with 72% accuracy on T-stage [8]. In addition, several Breast Imaging-Reporting and Data System (BI-RADS) classification approaches have been evaluated in English; the best results were obtained by using partial decision trees (PART) [9].

In Dutch, one study was published on free text mining in radiological reports using support vector machines (SVM) and conditional random fields (CRF) to structure free text data with a BI-RADS classification algorithm proposed as future work [10]. However, to our knowledge, no tumor-classification task based on radiology reports has been published in Dutch before.

This article describes a pilot study which shows the challenges to expect when extracting data from free text radiology reports in a small language area, like Dutch, in the classification of the T-stage of TNM pulmonary oncology.

Methods

Corpus description

After ethical approval at the participating medical center, a training set was created which consisted of 47 radiological reports with pulmonary oncology that underwent a diagnostic staging procedure. The radiological reports have been constructed by several different radiologists, other than the authors, using a speech recognition tool (G2 Speech). Findings were stored as free text reports in a Radiological Information System

(RIS, Agfa Healthcare). Every included report consisted of several structured sections with the following headings: clinical details, report, described modality, body part and conclusion. This training set was used to identify reporting content and to find appropriate synonyms, which were incorporated in the algorithm. Consecutively, a second set of 100 cases was used to validate the outcomes. Cases were included if a primary pulmonary malignancy was diagnosed using a computed tomography (CT) and the radiological report was present. Cases with two primary tumors and follow-up cases were excluded. After inclusion, T-stage was independently classified and labelled from the report by two authors (JMN and SP) according to the 8th TNM classification [2], because final T-stage was not explicitly mentioned in the report and could only be derived from findings described in the free text. The authors agreed on annotation guidelines for proper labeling. In case of discrepancy, consensus was reached between the two authors.

Algorithm structure

Because of the limited training data available, a rule-based NLP algorithm with machine learning pre-processing steps was used in this study. In addition, we aimed to set a baseline for future work using more advanced machine or deep learning techniques. The used approach is subdivided into a pre-processing step and a processing step. The pre-processing is necessary to make the data suitable for analysis. The processing step is the actual algorithm (Fig. 1, Table 1).

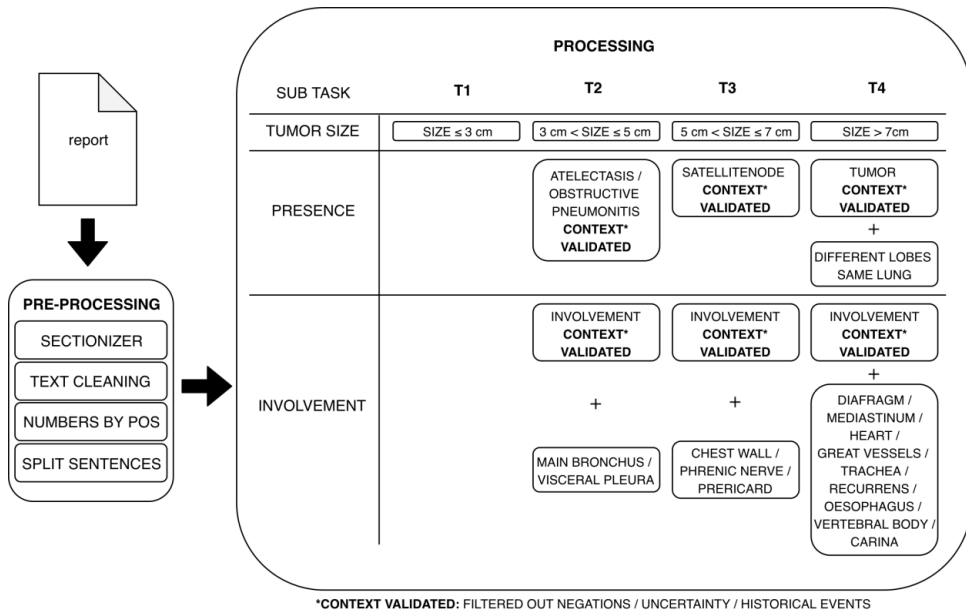


Figure 1. T-stage classifier. Schematic overview of T-stage classification. In the pre-processing step the raw data of the report is prepared for the actual processing. In the processing step tumor size extraction and a T-stage presence check of abnormalities and its involvement is performed.

Pre-processing

A sectionizer was developed to only select relevant parts of the report. In this study, text was only searched when preceded by the headings *thorax* and *conclusion*. A consecutive cleaning step was introduced to remove speech recognition artefacts and to replace selected abbreviations by its full form. Open-source NLP software library SpaCy [11] was selected to perform sentence segmentation and number extraction using part-of-speech tagging (POS), as it includes pre-trained models for multiple languages and has been successfully applied on medical extraction tasks before [12].

Processing

By analyzing the 8th TNM classification [2], the T-stage classification was divided into three different items: *size*, *presence* and *involvement* (Fig. 1). All three items required extraction of relevant concepts (e.g. tumor; Fig. 1). For every concept a set of synonyms and their conjugations was created (e.g. tumor; mass, lesion, etc.) to ensure a high recall

in extracting concepts from included reports. The synonym sets were created by radiological domain experts using the training set, Systematized Nomenclature of Medicine Clinical Terms (SNOMED-CT) [13] and their expertise.

Accordingly, the synonym sets were converted into a regular expression per concept. Depending on the item to extract (size, presence or involvement), the concepts were further processed by the algorithm in different ways.

To cover the item *size*, a measurement extractor was developed using POS recognition of NLP-library spaCy to extract tumor size. Tumor size was selected out of all numbers, when all of the following preconditions were fulfilled: the largest number, the number is part of an area expression, the number contains a unit (cm or mm), the number is not a distance measurement and is not preceded by the concept 'lymph node' (instead of 'tumor').

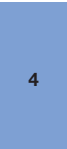
The concepts to extract for the item *presence* were context validated; for every extracted concept context information (for instance negations, uncertainty, historical events) was extracted. Only those concepts being certain by its related context were valid and used for classification.

pyContextNLP was used to extract the context including negations (modifier) related to the concept (target) , as it has been translated and applied to several languages, including Dutch [14][15][16]. pyContextNLP has been translated and functionality has been extended to run it as a service to simplify integration with other NLP services, increasing performance and usability [17].

Finally, to extract the item *involvement*, two different concepts had to be present in same sentence: the concept "involvement" itself and, the concept being involved (e.g. possible involvement in mediastinum). The concept "involvement" is context validated; context information (for instance negations, uncertainty, historical events) was extracted.

In addition, a specific T4-stage logic has been implemented to validate whether a tumor is present in different lobes of the same lung. Final T-stage was assigned to the most severe tumor classification found by the algorithm. A detailed example of the classification process is shown in Table 1: Detailed example of the classification process.

RAW REPORT	PROCESSED REPORT	CLASSIFIED REPORT								
<p>Clinical details: Pulmonary malignancy?</p> <p>Report: CT thorax and abdomen, arterial phase</p> <p>Thorax: Mass visible in the left upper lobe with a maximum size estimated at image 46 of 4, 7 x 3,0 cm. Possible involvement in mediastinum. Satellite nodes visible at 8-41 with an estimated size of 1,3 cm. Lymph node visible at station 7 with a size of circa 5,2 cm. No lymph nodes visible at contralateral side. Small consolidation middle lobe. No indication of atelectasis.</p> <p>Abdomen: Multiple sharply edged hypodense liver lesions visible which would initially match with cysts (HU 5).</p> <p>Musculoskeletal No relevant findings. No metastasis.</p> <p>Conclusion: Tumor with satellite nodes left upper lobe</p>	<p>Clinical details: Pulmonary malignancy?</p> <p>Report: CT thorax and abdomen, arterial phase</p> <p>Thorax: Mass visible in the left upper lobe with a maximum size estimated at image 46 of 4,7 x 3,0 cm. Possible involvement in mediastinum. Satellite nodes visible at 8-41 with an estimated size of 1,3 cm. Lymph node visible at station 7 with a size of circa 5,2 cm. No lymph nodes visible at contralateral side. Small consolidation middle lobe. No indication of atelectasis.</p> <p>Abdomen: Multiple sharply edged hypodense liver lesions visible which would initially match with cysts (HU 5).</p> <p>Musculoskeletal No relevant findings. No metastasis.</p> <p>Conclusion: Tumor with satellite nodes left upper lobe</p>	<table border="1" data-bbox="812 256 1111 496"> <tr> <td>Tumor size</td> <td>T1 (4,7 cm)</td> </tr> <tr> <td>Presence</td> <td>T3 (satellite nodes)</td> </tr> <tr> <td>Involvement</td> <td>-</td> </tr> <tr> <td>Classification</td> <td>T3</td> </tr> </table>	Tumor size	T1 (4,7 cm)	Presence	T3 (satellite nodes)	Involvement	-	Classification	T3
Tumor size	T1 (4,7 cm)									
Presence	T3 (satellite nodes)									
Involvement	-									
Classification	T3									



DESCRIPTION

Sectionizer: filtered out sections “Thorax” and “Conclusion”

Cleaning: Colons and whitespaces within numbers removed, selected abbreviations are replaced

Size: 4,7 cm is extracted as tumor size, the number is part of an area expression, has unit cm and is not preceded by lymph node.

Presence: pyContextNLP extracted concepts and **context**. "Mass" and "satellite node" is found without context.

Involvement: pyContextNLP extracted "involvement" with **context** of type uncertainty, therefore involvement in mediastinum is ignored.

Table 1. Detailed example of the classification process. Pre-processing is performed on the raw text of the report. In the processed report, only the relevant sections remain. Every sentence in the processed report is annotated with extracted measurements, concepts (presence/involvement) and context. The final classification is obtained by the highest T-stage detected.

Results

The accuracy of the T-stage classifier on the test set was 83% (N=47), and on the validation set 87% (N=100) (see Table 2: T-stage classifier accuracy).

	Training set (N=47)	Validation set (N=100)
Accuracy T-stage	0.83	0.87

Table 2. T-stage classifier accuracy scores of the training set and the validation sets.

Figure 2 shows the confusion matrices of respectively the training set and the validation set, where each ‘actual T-stage’ is compared with the ‘predicted T-stage’.

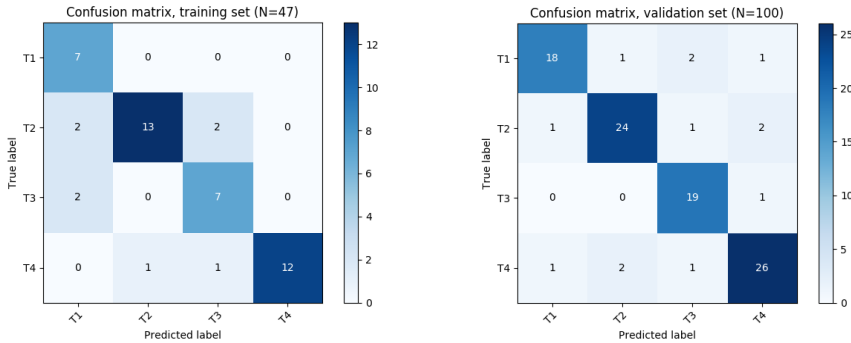


Figure 2. Confusion matrices of the T-stage classification on the training and validation sets.

The precision (i.e. specificity), recall (i.e. sensitivity) and F₁ measure (i.e. combined metric for precision and recall) for all independent stages are obtained as shown in Table 3: Precision, recall and F₁-scores.

Training	Precision	Recall	F ₁ score
T1	0,64	1,00	0,78
T2	0,93	0,76	0,84
T3	0,70	0,78	0,74
T4	1,00	0,86	0,92

Validation	Precision	Recall	F ₁ score
T1	0,90	0,82	0,86
T2	0,89	0,86	0,87
T3	0,83	0,95	0,88
T4	0,87	0,86	0,87

Table 3. Precision, recall and F₁-scores of training and validation sets.

In addition, all errors in the training set and validation set were analyzed and grouped into five specific categories with one or more subgroups: context, concepts, standardization, complexity, and spaCy (Table 4). In total seven errors were found in the training set and 13 in the validation set.

Finally, in Appendix 1 (Concept synonyms) SNOMED concepts have been added to the table of used regular expressions, to indicate the number of translations and synonyms missing. In Appendix 2 (Mentions related to context) and Appendix 3 (Mentions related to involvement) challenges related to context and involvement are highlighted to point out difficulties of the process.

Error Group	Error Type	Description	Training (n=47)	Validation (n=100)
Context	Context missing	Context not matched because of missing modifier	0	1
	Context mismatch	Context mismatch, wrong modifier detected	2	3
	Context disagreement	Disagreement about certain/prob certain	0	1
Concepts	Missing synonym	Concept not matched because of a missing synonym or expression.	2	0
	Algorithm logic	Presence or involvement not correctly classified	0	2
Standardization	Measurement Extractor	e.g. using expressions (more than 5 cm) or 4-51 op 11 cm, blacklist for size	2	2
	Dictation Artifact	Errors related to dictation (e.g. whitespaces within numbers)	0	1
	Standardization	Wrong heading above section	0	1
Complexity	T4 multiple lobes	Error related by detecting tumor present in multiple lobes of the same lung	1	1
spaCy	Sentence Boundary Detection	Error in detecting the boundary of a sentence, therefore involvement logic does not hold	0	1
	Total Errors		7	13

Table 4. T-stage errors by category for the training and validation sets.

Discussion

The aim of this paper is to gain insight in the challenges of using NLP in free text radiological reports in a small language area such as Dutch. This was done by creating an algorithm for T-stage pulmonary oncology according to the 8th TNM classification. This feasibility study is a baseline for future work based on more (hybrid) advanced machine or deep learning techniques.

The described method analyzes and tries to thoroughly understand the meaning and interactions of words and phrases in the radiological report before classifying it. The main difference with a general machine or deep learning approach is that different steps are used before the final analysis is performed, instead of analyzing the report as a whole. Because the TNM classification is already rule-based, it is not necessary to force the neural network to recompose the already known T-stage rules for proper T-staging. Focusing on how to properly analyze free text was therefore one of the main goals of this approach as this can show us where difficulties can be expected and where machine or deep learning can help us smoothen this process.

The measured accuracy of this pilot study suggests that T-stage can be extracted from free text reports with a fairly high reliability. This is consistent with the earlier performed study on pathology reports written in English [8]. In addition, the strategy used for extracting the items *size*, *presence* and *involvement* according to the 8th TNM classification seems promising. The obtained results (precision, recall and F₁ score) for the training and validation set are in most cases at least comparable.

When looking at the pre-processing and processing steps, several important findings should be addressed. First of all, identification of synonyms of the chosen items is of utmost importance, because vocabulary used for describing tumors differs widely among reporters. This variability in vocabulary makes it difficult to use machine learning for finding appropriate synonyms at this stage, because a large amount of data is needed. However, when a sufficient amount of data is available word embeddings could be created, which might be used to automatically find synonyms for used concepts. This study highlights the importance of using domain specific knowledge when building a (rule-based) algorithm when training data is limited.

Attempts to find proper synonyms by using (the Dutch) SNOMED-CT failed. Used synonyms are not always a synonym of the proper SNOMED-CT concept, but for example a synonym of a related super concept. Iterating over all supertype (parent) concepts is tedious and most are irrelevant (e.g. several tumor synonyms can be found searching for abnormal morphology). In addition, the Radiological Lexicon (RadLex) was not available in Dutch and could therefore not be tested. Ideally, a standardized vocabulary should be used to standardize data and try to make data more uniform. Data should then be labelled with SNOMED-CT or RadLex codes in order to increase findability, according to the Findable, Accessible, Interoperable and Reusable (FAIR) principles [18].

Another important finding is that radiological free text reports consist of many contextual expressions, phrases and words (see Appendices 2 and 3), which are indispensable for accurate description of a specific disease. For instance, concepts should be properly correlated to the right context like negations or sizes, but the same holds for probabilities and the extent of involvement. This is a difficult and important process and should be done with care, because context allows radiologists to nuance and specify their findings. This lack of nuancing possibilities is probably one of the caveats of structured reporting and its broad implementation.

When analyzing the errors in detail, one can see that the errors made are diverse, although most wrongly staged tumors were related to context extraction (35%). Several times there is a mismatch between concept and context caused by the shallow approach of pyContextNLP. For example, when two concepts are present in the same sentence, context (e.g. a negation) can be matched with the wrong concept. This might be overcome by dependency parsing which can improve contextual matching.

This paper tried to divide pre-processing and processing steps in order to differentiate errors found, but the errors are often hard to separate, as both steps are highly intertwined. For instance, errors made by the sentence splitter can be related to the fact that the model is not trained on medical reports. However, errors can also be introduced by radiological reporters using a different (staccato) way of reporting. The use of speech recognition in radiological reporting introduces several imperfections, mainly resulting in incorrect punctuation and white space errors within numbers. This can only be partly improved by pre-processing steps.

Task complexity is a different hurdle to overcome. Problems might, for instance, arise when concepts of different items should be combined in a single statement (e.g. T4-stage, different lobes, same lung) or should be ignored (e.g. gravity depending atelectasis vs. tumor related atelectasis). This is especially the case when these concepts are stated in different sentences. Specific annotation guidelines or agreements can partly improve this difficulty. However, algorithms should not be unnecessarily more complicated when steps like standardization of the report content or reporting manner can increase report homogeneity. This is highlighted by the errors made in the standardization category (30%) which are related to the input of the reporter and dictation technology used. Standardizing report content by using a certain standardized language, for instance the vocabulary used in the TNM classification, will result in less synonyms in the report. In addition, when sentences stated are less ambiguous, by for instance stating only information about the described item in the same sentence, outcomes will further be improved. As such, standardization of reporting content and manner will improve outcomes without expanding existing algorithms. Hence, NLP and standardization are counterparts in which high-end NLP tooling makes standardization redundant, but proper standardization can improve the structured data and the accuracy of the NLP tool.

Several limitations of this study should be mentioned of which the small sample size is the most important one. Furthermore, this algorithm is only trained at one specific dataset of one radiological department. Therefore, overfitting is a concern. Although this has not been the main goal of this pilot, future work should focus on external validation.

In addition, future work should be done to explore how NLP algorithms can increase the value of the radiological report when, for instance, they are incorporated in the reporting process. Live classifications can be displayed when an algorithm is processing the free text during reporting. An algorithm can also notify the reporter when information about a specific item is missing. In addition, this tumor staging algorithm can also be used for restaging earlier staged tumors according to the current TNM edition. As such, NLP algorithms can be used in various ways to enhance reporting content and support the FAIR principles.

Conclusion

NLP is a promising technology for mining free text radiological reports and can be introduced in English and in a small, non-English language such as Dutch. However, the proper implementation of a free text algorithm depends largely on the context of concepts mentioned in the report, more than on specific words. Implementing NLP and standardization should be balanced, and ratios adjusted depending on the available data. Future work should mainly focus on how to (gradually) use more machine or deep learning approaches.

References

1. McGinty GB, Allen B, Geis JR, Wald C. IT infrastructure in the era of imaging 3.0. *J Am Coll Radiol*. 2014;11(12 Pt B):197-204. doi: 10.1016/j.jacr.2014.09.005.
2. Brierley J, Gospodarowicz MK, Wittekind C, editors. *TNM classification of malignant tumours*. 8th ed. Chichester: John Wiley & Sons Inc; 2017.
3. Puts S, Nobel JM. Medical narrative to structure: maastroclinic/medstruct. maastroclinic; 2019.
4. Krupinski EA, Hall ET, Jaw S, Reiner B, Siegel E. Influence of radiology report format on reading time and comprehension. *J Digit Imaging*. 2012;25:63-69 doi: 10.1007/s10278-011-9424-8.
5. Pons E, Braun LMM, Hunink MGM, Kors JA. Natural Language Processing in Radiology: A Systematic Review. *Radiology*. 2016;279(2):329-43. doi: 10.1148/radiol.16142770.
6. Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *J Am Med Inform Assoc*. 2010;17(5):507-13. doi: 10.1136/jamia.2009.001560.
7. Cornet R, van Eldik A, de Keizer N: Inventory of Tools for Dutch Clinical Language Processing. *Stud Health Technol Inform*. 2012;180:245-9.
8. Nguyen AN, Lawley MJ, Hansen DP, Bowman RV, Clarke BE, Duhig EE, et al. Symbolic rule-based classification of lung cancer stages from free-text pathology reports. *J Am Med Inform Assoc*. 2010;17(4):440-5. doi: 10.1136/jamia.2010.003707.
9. Castro SM, Tseytlin E, Medvedeva O, Mitchell K, Visweswaran S, Bekhuis T, et al. Automated annotation and classification of BI-RADS assessment from radiology reports. *J Biomed Inform*. 2017;69:177-187. doi: 10.1016/j.jbi.2017.04.011.
10. Pathak S, van Rossen J, Vijlbrief O, Geerdink J, Seifert C, van Keulen M. Post-Structuring Radiology Reports of Breast Cancer Patients for Clinical Quality Assurance. *IEEE/ACM Trans Comput Biol Bioinform*. 2020;17(6):1883-1894. doi: 10.1109/TCBB.2019.2914678.

11. Honnibal M, Montani I. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. To appear: 2017.
12. Soldaini L, Goharian N. QuickUMLS: a fast, unsupervised approach for medical concept extraction. MedIR workshop [Internet]. 2016 [cited 6 May 2019]. Available from <http://ir.cs.georgetown.edu/downloads/quickumls.pdf>.
13. Côté RA, Robboy S. Progress in medical information management. Systematized nomenclature of medicine (SNOMED). JAMA. 1980;243(8):756-62. doi: 10.1001/jama.1980.03300340032015.
14. Chapman BE, Lee S, Kang HP, Chapman WW. Document-level classification of CT pulmonary angiography reports based on an extension of the ConText algorithm. J Biomed Inform. 2011;44(5):728-37. doi: 10.1016/j.jbi.2011.03.011.
15. Chapman WW, Hillert D, Velupillai S, Kvist M, Skeppstedt M, Chapman BE, et al. Extending the NegEx Lexicon for Multiple Languages. Stud Health Technol Inform. 2013;192:677-681.
16. Afzal Z, Pons E, Kang N, Sturkenboom MC, Schuemie MJ, Kors JA. ContextD: an algorithm to identify contextual properties of medical terms in a Dutch clinical corpus. BMC Bioinformatics. 2014;15(1):373. doi: 10.1186/s12859-014-0373-3.
17. Chapman WW. Extract context modifiers targeting clinical terms: maastroclinic/pyConTextNLP [Internet]. Maastricht: maastroclinic; 2019. Available [cited 19 June 2019]. Available from <https://github.com/maastroclinic/pyConTextNLP>.
18. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016;3:160018. doi: 10.1038/sdata.2016.18. Erratum in: Sci Data. 2019;6(1):6.

Supplementary material

Appendix 1: Concept synonyms

Concept	Regular expression (Dutch)	SNOMED-CT concept (description in Dutch)	SNOMED-CT subset of relevant supertype - subtype relationships (Dutch)
tumor	"(massa tumor nodu haard carci no laesie letsel \brip\b \bRIP\b r uimte(-)?innemend process maligniteit verdicht indicht)"	108369006 neoplasma (afwijkende morfologie)	4147007 massa 27925004 nodus 52988006 afwijkend weefsel 19130008 traumatisch letsel 9656002 consolidatie 707496003 inflammatie en consolidatie 367651003 maligne neoplasma van primaire, secundaire of onzekere oorsprong
involvement	"aan(ge)?tast destructie (door in)(ge)?groei uitbreiding betrokken invade invasie induratie"	248448006 Involved	66211004 Extending 385394007 Tumor invasion by site 129382001 destructie 45147008 induratie
lymph nodes	lymf nodes klier	59441001 structuur van nodus lymphaticus	

T2 Presence

main_bronchus	"(centrale hoofd hilair).*bronch	102297006 structuur van bronchus principalis	
visceral_pleura	pleura longvlies"	81623005 structuur van pleura pulmonalis	

T2 Involvement

atelectasis	atelect samengevallen	46621007 atelectase (aandoening)	
obstructive_pneumonitis	obstructieve pneumoni infect.*verander	205237003 pneumonitis (aandoening)	

T3 Involvement

chest_wall	borst.*wan thorax.*wan rib costa "	78904004 structuur van thoraxwand (lichaamsstructuur)	113197003 botstructuur van costa (lichaamsstructuur)
nervus_phrenicus	nervus.*(phrenicus frenicus)"	50230006 structuur van nervus phrenicus (lichaamsstructuur)	
parietale_pericard	pericard	76848001 structuur van pericardium (lichaamsstructuur)	

T₃ Presence

satellite_nodule	satelliet.**	396408009 Specimen involvement by satellite nodule(s) present (finding)	
------------------	--------------	---	--

T₄ Involvement

diaphragm	diafragm middenrif	5798000 structuur van diaphragma (lichaamsstructuur)	
mediastinum	mediast	72410000 structuur van mediastinum (lichaamsstructuur)	427352001 Tumor invades mediastinum (finding)
heart	hart epicard	80891009 structuur van cor (lichaamsstructuur)	6871001 structuur van epicardium (lichaamsstructuur)
great vessels	grote vaten centrale vaten aorta vena cava vcs VCS	3711007 structuur van grote vaten (lichaamsstructuur)	81040000 structuur van arteria pulmonalis (lichaamsstructuur) 3711007 structuur van grote vaten (lichaamsstructuur) 181368006 gehele vena cava superior (lichaamsstructuur) 15825003 structuur van aorta (lichaamsstructuur)
trachea	trachea luchtpijp	44567001 structuur van trachea (lichaamsstructuur)	
recurrent_laryngeal_nerve	recurrens	731050007 gehele nervus laryngeus recurrens (lichaamsstructuur)	280300006 structuur van linker nervus laryngeus recurrens (lichaamsstructuur) 280299003 structuur van rechter nervus laryngeus recurrens (lichaamsstructuur)
oesophagus	oesophagus slok oesofagus oesof oesop	32849002 structuur van oesofagus (lichaamsstructuur)	
vertebral body	wervel vertebra	3572006 structuur van corpus vertebrae (lichaamsstructuur)	420345000 structuur van vertebra (lichaamsstructuur)
carina	carina	28700002 structuur van carina trachea (lichaamsstructuur)	

T₄ Tumor in different lobes

superior_lobe_right	echter\s{o,i}bovenkwab boven\s{o,i}kwab rechts \bRBK\b \bRBL\b	362898004 structuur van kwab van rechter long (lichaamsstructuur)	
superior_lobe_left	linker\s{o,i}bovenkwab boven\s{o,i}kwab links \bLBK\b \bLBL\b	44714003 structuur van bovenkwab van linker long (lichaamsstructuur)	
middle_lobe	midde.*(kwab lob) \bMK\b \bML\b	72481006 structuur van lobus medius pulmonis dextri (lichaamsstructuur)	
inferior_lobe_right	"rechter\s{o,i}onder\s{o,i}(kwab lob) onder\s{o,i}(kwab lob)recht \bROK\b \bROL\b"	266005 structuur van onderkwab van rechter long (lichaamsstructuur)	

inferior_lobe_left	"linker\{o,i}onder\{o,i}(kwab lob) onder\{o,i}(kwab lob)link \bLOK\b \bLOL\b"	41224006 structuur van onderkwab van linker long (lichaamsstructuur)	
--------------------	---	--	--

Regular expressions used for key classification concepts tumor and involvement, corresponding SNOMED-CT concepts and incomplete list of relevant supertype-subtype relationships can be found below. In the column 'Regular expression' the colors indicate if the synonym could be found (green) or could not be found (red) in the Dutch edition of SNOMED-CT.

Appendix 2: Mentions related to context

Uncertainty mentions found in reports (context categories ignored for involvement and present concepts)	
Original mention in Dutch	Translation in English
dit kan een ... cave enige massawerking ter plaatse van de vaten / op de vaten zichtbaar lijkt Ingroei in het mediastinum op basis van deze scan niet duidelijk zichtbaar. een en ander verdacht voor primair carcinoom zonder aanwijzingen voor ingroei "Doorgroei in het ... is niet met zekerheid uit te sluiten. " "Ook hier mogelijk enige uitbreiding buiten het longvlies." geen directe ingroei zichtbaar is het beeld niet geheel typisch voor primair longcarcinoom. mediastinale ingroei niet geheel is uitgesloten verdenking op satellietlaesie	this can be a... be aware of some mass effect at the location of the vessels / on the vessels seems visible Extension into the mediastinum is not clearly visible , based on this scan. suspected for primary carcinoma without indication of involvement "Extension in the ... cannot be completely ruled out. " "Here too, extension outside the peritoneum is possible. " no direct ingrowth visible the imaging picture is not typical for primary lung carcinoma. mediastinal involvement cannot fully be excluded suspicion of satellite lesion

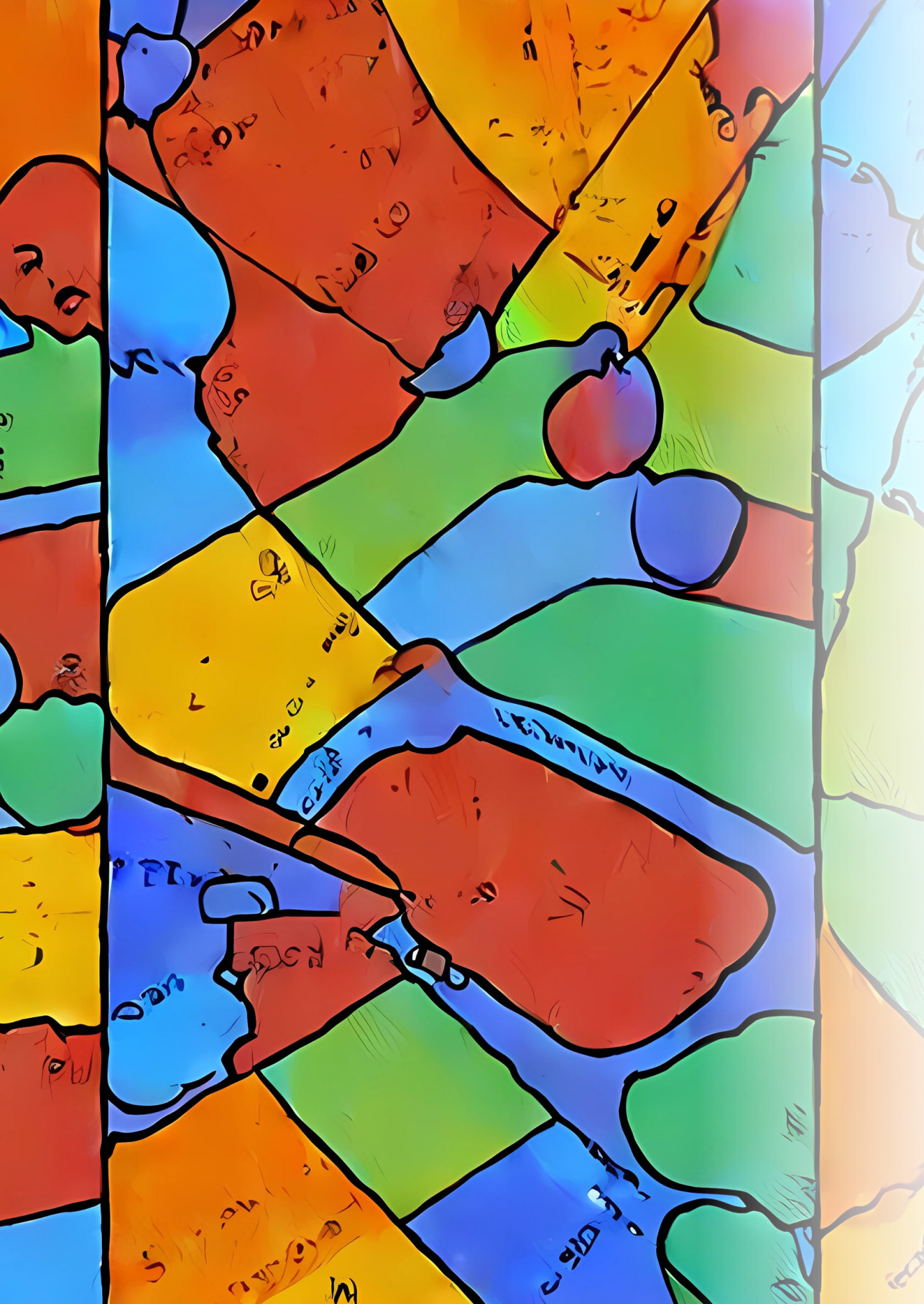
These are actual mentions selected from the reports in the corpus to highlight the extensive usage of context in the language used.

Appendix 3: Mentions related to involvement

Involvement mentions (not incorporated in the algorithm)	
Original mention in Dutch	Translation in English
afgrenzing tegen liggen tegenaan contact makend met	boundary to lay towards towards making contact with close contact

nauw contact	some mass effect
enige massawerking	very close relationship
zeer nauwe relatie	spur reaching
uitloper reikend tot tegen	spur towards
uitlopers naar	broad basic contact
breedbasisch contact	visible towards
zichtbaar tegen	partly towards ... adjacent to
deels tegen ... aangelegen	mediastinum compression
mediastinum compressie	spur towards the
uitloper richting de	traction on the pleura
tractie aan de pleura	irregular spicular boundaries and shows traction on ...
onregelmatige spiculaire begrenzingen en vertoont tractie ...	the mass reaches towards ...
hierbij reikt de massa tot aan ...	mass extends to the carina
massa breidt zich uit tot aan de carina	be aware of ingrowth
cave ingroei	without signs of cortical destruction
zonder dat er tekenen zijn van een corticale aantasting	there is a close relationship with the aorta, but without
er is een nauwe relatie met de aorta doch zonder overtuigende	convincing signs of ingrowth
aanwijzingen voor ingroei	signs of ingrowth
tekenen van ingroei	

These mentions are not incorporated in the algorithm (e.g. regular expressions or 'regex' for involvement) but show the usage of free text to indicate almost involvement.





Chapter 5:

T-staging pulmonary oncology from radiological reports using Natural Language Processing:
translating into a multi-language setting

*J. Martijn Nobel, Sander Puts, Jakob Weiss, Hugo J.W.L. Aerts,
Raymond H. Mak, Simon G.F. Robben, André L.A.J. Dekker*

Insights into Imaging (2021)

Abstract

Background: In the era of datafication it is important that medical data is accurate and structured for multiple applications. Especially data for oncological staging needs to be accurate to stage and treat a patient, as well as population-level surveillance and outcome assessment. To support data extraction from free text radiological reports, a Dutch Natural Language Processing (NLP) algorithm was built to quantify T-stage of pulmonary tumors according to the Tumor Node Metastasis (TNM) classification. This structuring tool was translated and validated on English radiological free text reports. A rule-based algorithm to classify T-stage was trained and validated on respectively 200 and 225 English free text radiological reports from diagnostic Computed Tomography (CT) obtained for staging of patients with lung cancer. The automated T-stage extracted by the algorithm from the report was compared to manual staging. A Graphical User Interface was built for training purposes to visualize the results of the algorithm by highlighting the extracted concepts and its modifying context.

Results: Accuracy of the T-stage classifier was 0.89 in the validation set, 0.84 when considering the T-substages, and 0.76 when only considering tumor size. Results were comparable with the Dutch results (respectively 0.88, 0.89 and 0.79). Most errors were made due to ambiguity issues that could not be solved by the rule-based nature of the algorithm.

Conclusions: NLP can be successfully applied for staging lung cancer from free text radiological reports in different languages. Focused introduction of machine learning should be introduced in a hybrid approach to improve performance.

Introduction

Radiological reports contain an extensive amount of historical information about the patient and their current disease status over a prolonged period of time [1]. Ideally, information from such reports should be available as structured data that can easily be communicated and reused. Instead, these reports are generally at best semi-structured free text reports, which takes a human reader to interpret. Natural Language Processing (NLP) techniques provide solutions for the extraction of structured data from

unstructured text and has been applied to many healthcare purposes and may help to extract structured information from radiology reports [2].

Specific NLP algorithms already exist to find tumor specific information in radiological reports to extract, for instance, cancer outcomes [3-6]. Next to extracting tumor endpoints and follow-up from radiological reports, NLP algorithms can also be used to extract tumor staging from free text. An example is a Dutch rule-based NLP algorithm that can extract the T-stage for lung cancer according to the Tumor Node Metastasis (TNM) oncology classification system from the free text radiological reports of chest Computed Tomography (CT) scans [7, 8]. Lung cancer is the most common oncological cause of death, with imaging playing a great part in its diagnosis and staging [9]. Therefore, improvements in the reporting and staging process may be valuable. Specifically, it may speed up workflow and enhance the quality and accuracy of the radiological report, as well as communication between health professionals.

The Dutch algorithm analyses the radiological report and extracts tumor stage with an accuracy score of 0.83-0.87. In addition, this algorithm can also be used for (re)staging historical data, which may be useful, for instance, in cases that have been classified with an older version of the TNM classification system or adjustments with newly available data. An NLP algorithm can therefore function as an important solution to increase the value of the radiological report. Implementation of this NLP algorithm can also act as a method to extract and convert unstructured free text information into stored structured information from radiological reports. This is important, because structured stored data can be processed more easily than free text for clinical or research purposes [10]. This is of particular interest when realizing that over the past years a shift towards structured reporting has been promoted by the Radiological Society of North America (RSNA) and the European Society of Radiology (ESR). The goal of this is to increase the value of the radiological report and allow for better content datafication [11, 12]. Moreover, the ESR published guidelines for radiologists on reporting and good practice, which highlights the need for better reporting, also promoting the potential of (multilingual) structured reporting [13, 14]. Also, several surveys of radiologists show a global shift towards the use of structured reporting in radiology [15, 16], as many radiologists appreciate the benefits of structured reporting, such as report clarity, communication and data mineability [17, 18]. Although the NLP approach does not use a strict structured

reporting format like a point-and-click system, drop down menu or template to insert structured data elements, it does analyze the old-fashioned free text report to create structured data during or after the reporting process. Thereby, this NLP algorithm can also be used on old free text reports to extract T-stage according to the current standards and can help quality assessments of oncological registries.

This algorithm is only capable of processing Dutch staging CT reports and is therefore only proven to be effective in Dutch. With the translation of used regular expressions it may be possible to translate the algorithm into other languages, like English. In addition, to increase understanding of the algorithm and to utilize its full potential, building a Graphical User Interface (GUI) might increase the usability and clinical utility of the algorithm. The hypothesis of this study is that the Dutch algorithm can be translated into English to allow for analysis of English free text radiological reports.

This paper presents the process of translation, implementation and validation of the Dutch pulmonary T-stage algorithm to reports written in English with the use of a GUI.

Methods

Corpus description

After institutional review board approval at the participating medical center, an existing retrospective lung cancer clinical database of patients treated at the institution was used to search for radiological reports of diagnostic CT or Positron Emission Tomography-Computed Tomography (PET-CT) scans, performed at initial cancer staging. Inclusion totaled 425 radiological reports of patients with primary pulmonary oncology of which the full report of the staging examination was available. Cases were excluded in case of 1) follow-up and restaging reports (second opinions), 2) cases with two primary tumors or 3) incomplete reports (no full text and/or primary staging report available). The first 200 reports formed a training set, the remainder of the cases composed a validation set (n = 225). Tumor and report characteristics of both groups are shown in Table 1.

	Training (n = 200)	Validation (n = 225)
TNM substage		
<i>T1a</i>	4	6
<i>T1b</i>	27	31
<i>T1c</i>	42	42
<i>T2</i>	6	3
<i>T2a</i>	32	44
<i>T2b</i>	27	23
<i>T3</i>	33	41
<i>T4</i>	29	35
Report format		
<i>CT</i>	106	120
<i>PET</i>	77	88
<i>PET/CT</i>	17	17

Table 1. Cohort composition of the training and validation sets Included report statistics by T-substage for the training and validation sets.

Determining T-stage

The radiology reports were created using a speech recognition device and contained free text concerning at least the lungs. Three different report formats that could be discerned were all included: a strictly radiological CT report, a PET-CT report in which radiological information was blended with the nuclear diagnostic information and a more structured PET-CT format in which the two types of information were separated in the report. Most reports used subheadings for the body part lung, like *Thorax* or *Chest*. Also, other body parts were described in most of the reports and consisted of different combinations of the following elements: *History, Comparison, Technique, Findings (CT and/or PET-CT), Head, Neck, Chest, Mediastinum, Abdomen, Pelvis, Bones and Musculoskeletal*.

Because TNM-stage was not separately mentioned in these clinical reports, the T-stage was classified manually retrospectively from the report, according to the AJCC 8th edition TNM classification [7]. The authors agreed on annotation guidelines for proper

labeling and the T-stage was only scored if it was stated as being certain. In ambiguous cases, final T-stage was determined after reaching consensus between two authors.

Modifications for use in English

The training set was used to identify the specific structure and the indentation of the reports. Furthermore, the used subheadings had to be identified in the training set to correctly whitelist or blacklist specific sections of the report. To find proper English synonyms, the Dutch regular expressions, containing all synonyms and variants which are linked to the Systematized Nomenclature of human MEDicine-Computed Tomography (SNOMED-CT) terms [19] were translated and used as a starting point. These Dutch regular expressions were used to build an English Regular Expression (RegEx) per concept, which included the accompanying SNOMED-CT label to assure for proper ontology-based standardized classification. The used synonyms in English and their accompanying RegEx and ontology-based SNOMED-CT terms can be found in Appendix 2.

Algorithm structure

This study used the same lung cancer T-stage algorithm structure as the Dutch language-based algorithm, in which processing was subdivided in a preprocessing and a processing step to consecutively clean and process the radiological report [8]. Three similar items from the T-staging method had to be extracted (*size, presence and involvement*) before the T-stage classifier was able to stage the full T-stage (Fig. 1). Open-source part-of-speech (POS) tagging, NLP software library spaCy and pyContextNLP were used for number extraction, sentence segmentation and context validation [20, 21]. In addition to the Dutch algorithm, a blacklist had to be added to ignore sentences containing (mass) sizes in organs or body parts other than the lung, as some PET-CT scans covered more than only the thorax.

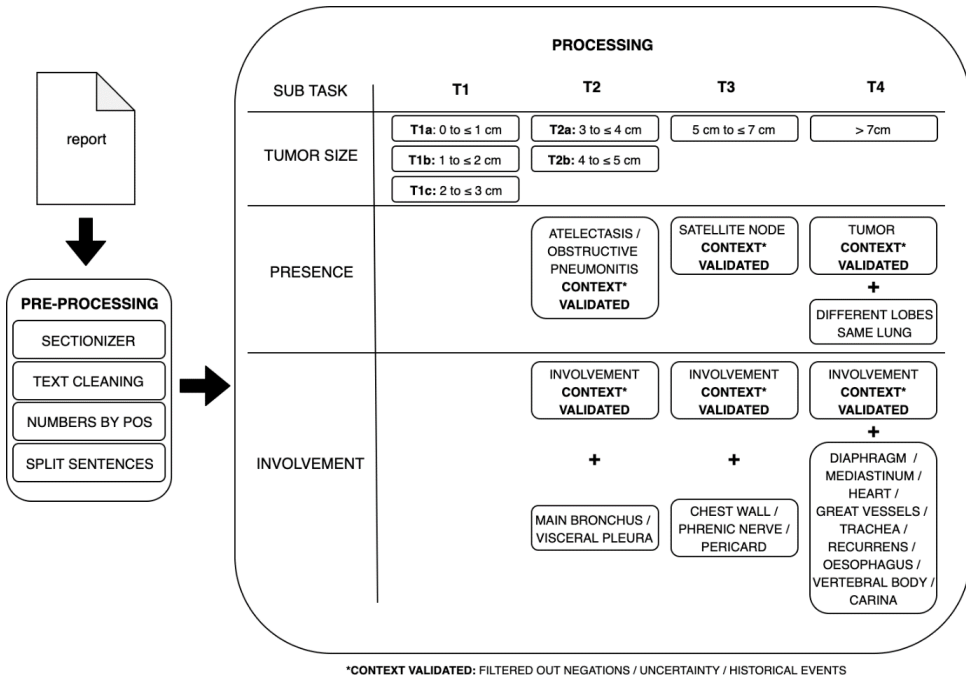


Figure 1. T-stage classifier. Schematic overview of T-stage classification. In the pre-processing step the raw data of the report is prepared for the actual processing. In the processing step tumor size extraction and a T-stage presence check of abnormalities and its involvement is performed.

Analysis

Analysis of the data was performed in order to assess the separate accuracy scores of the training and validation set for the T-stage (e.g. T1-T4) and the T-substage (e.g. T1a, T1b, T1c). In addition, T-stage, in which only size was used for classification, was calculated and compared with the Dutch results. Recall (i.e. sensitivity), precision (i.e. specificity), and F₁ measure (i.e. combined metric for precision and recall) for the T-stage classifier have been calculated for all substages in the training and validation set. To further differentiate outcomes, the total number of errors were grouped by category into context, concepts, standardization, complexity ambiguity, preprocessing and reporter.

Graphical User Interface

For this study a GUI, called MedStruct, was built to train and visualize the results of the algorithm by highlighting the extracted concepts and its modifying context (Fig. 2) [2].

MEDSTRUCT-NLP

Assistance, Classification and Information Extraction for Medical Free-Text Reporting

View Annotated:

Autocheck:

← →

Clinical details:
Pulmonary malignancy?

Report:
CT thorax and abdomen, arterial phase

Thorax:
Mass visible in the left upper lobe with a maximum size estimated at 8-46 of 4.7 x 3.0 cm. Possible involvement in mediastinum. Satellite nodes visible at 8-41 with an estimated size of 1.3 cm. Lymph node visible at station 7 with a size of circa 5,2 cm. No lymph nodes visible at contralateral side. Small consolidation middle lobe. No indication of atelectasis.

Abdomen:
Multiple sharply edged hypodens lesions visible which would initially match with cysts (HU 5). No relevant musculoskeletal findings

Conclusion:
Tumour with satellite nodes left upper lobe

POWERED BY TINY

Example text from Article-1 (EN)

TNM-8 Lung

English

T3

Primary Tumor

4.7 cm

Satellite Nodule (T3) Ipsilateral Tumor (T4)

Lymph Nodes

MEDSTRUCT-NLP

Assistance, Classification and Information Extraction for Medical Free-Text Reporting

View Annotated:

Autocheck:

Annotated Report

ThoraxMass visible in the left upper lobe with a maximum size estimated at 8 46 of 4.7 x 3.0 cm. Possible involvement in mediastinum. Satellite nodes visible at 8 41 with an estimated size of 1.3 cm. Lymph node visible at station 7 with a size of circa 5,2 cm. No lymph nodes visible at contralateral side. Small consolidation middle lobe. No indication of atelectasis.

- Present
- Involved
- Context Modifier
- Context Target

Example text from Article-1 (EN)

Satellite Nodule (T3) Ipsilateral Tumor (T4)

Lymph Nodes

Submit Feedback

TNM edited Report edited

Figure 2. Two screenshots of the Graphical User Interface MedStruct with the original report on the left side and its T-stage on the right side, combined with the items size, present and involvement. Also N (nodal stage) and M (metastatic disease) are mentioned for future use. By using drop down menus stages can be adjusted (upper). Annotated report at the left side and a feedback form at the right (lower). [22]

This was especially useful for finding proper synonyms as well as for analyzing and adjusting errors during training. To enable this GUI, the algorithm has been re-implemented into five reusable NLP pipeline microservices without changing the approach of the algorithm nor the algorithm itself (Fig. 3). The total pipeline now consists of a preprocessing component, spaCy, pyContextNLP, measurement extractor and the T-stage classifier. A web application has been created in which the report can be inserted or edited. The T-stage classification is automatically extracted and the result is immediately displayed.

The GUI highlights concepts and modifiers found in the report and displays the location, size, presence and involvement items on which the T-stage is based. Items can be adjusted using implemented drop down menus.

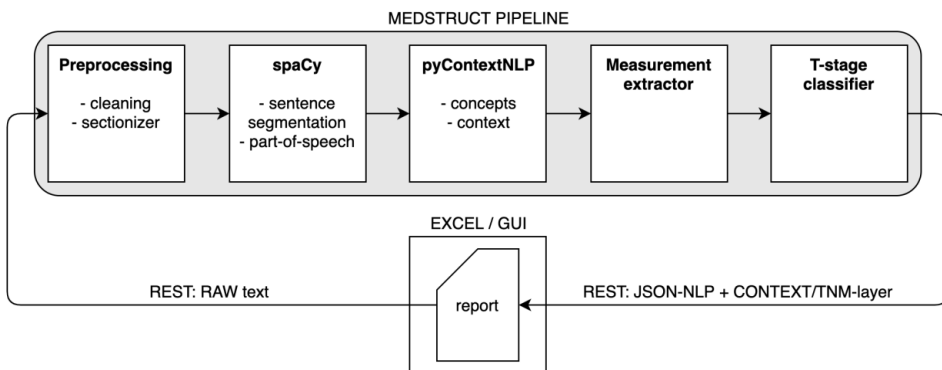


Figure 3. Schematic overview of the MedStruct pipeline, in which five different microservices are present: preprocessing, spaCy, pyContextNLP, measurement extractor and T-stage classifier. The report can be processed either from an Excel file or direct from the Graphical User Interface (GUI). All components use an intermediate JavaScript Object Notation (JSON) annotation format to chain the pipeline components and can be consumed over REpresentational State Transfer (REST) or chained using a message broker. The use of a JSON annotation format simplifies reusability of the different components, enables mixing programming languages, prevents for duplicate processing and guarantees token alignment between components. This implementation saves annotations at token level instead of sentence level, which enables precise highlighting of annotations in a GUI. Detected tumor and lymph nodes are stored as objects in a list, allowing for detection of concurrent mentions. Documents can now be processed individually with the same rule-based algorithm.

Results

Algorithm performance

The manually annotated T-stages and the report formats of the included reports were equally distributed in the training and validation set (Table 1). Only substages T1a and T2 have lower F₁-scores in both the training and validation set compared to the other stages. This might be due to the fact that these are underrepresented in this cohort.

The T-stage classifier accuracy was 0.89 for both the training and validation set, 0.87 and 0.84 when considering the T-substages and 0.78 and 0.76 when only using tumor size for classification (Table 2).

	English		Dutch	
	Training (n = 200)	Validation (n = 225)	Training (n = 47)	Validation (n = 100)
Accuracy T-substage	0.87	0.84	0.79	0.88
Accuracy T-stage	0.89	0.89	0.81	0.89
Tumor size based T-stage	0.78	0.76	0.70	0.79

Table 2. T-stage classifier accuracy scores of training set and validation sets in the English cohort and the Dutch cohort. In the Dutch group, the outcomes with the new processing structure are recalculated at the substage level.

The accuracy rates of the Dutch algorithm are added in the same table, showing the same outcomes in respectively the training (n = 47) and validation set (n = 100). A confusion matrix is shown in Fig. 4, where actual T-stage (*true label*) is compared with the predicted T-stage (*predicted label*).

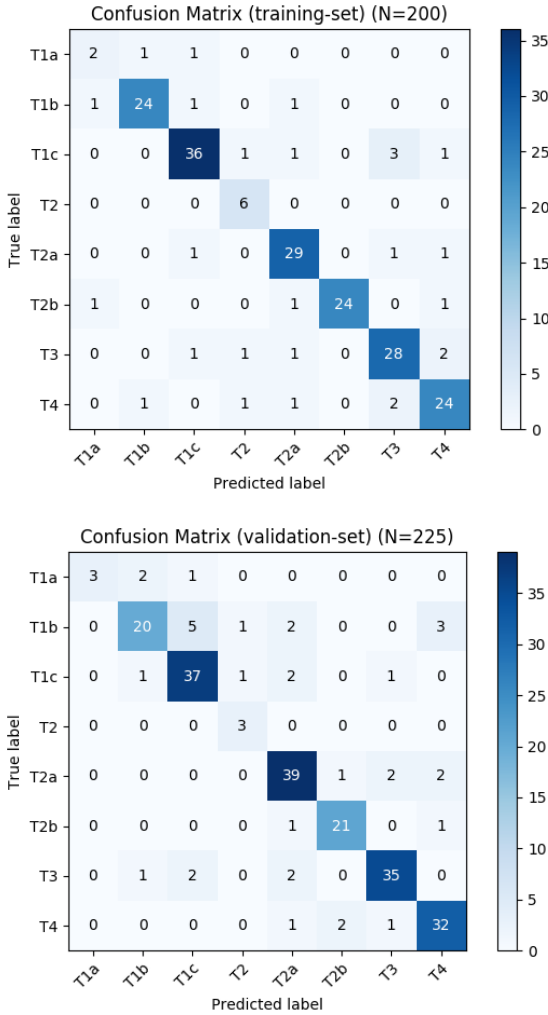


Figure 4. Confusion matrices of the T-stage classification on training and validation sets. Confusion matrices of the T-stage classification training set (upper) and validation set (lower)

In addition, the recall (i.e. sensitivity), precision (i.e. specificity), and F_1 measure (i.e. combined metric for precision and recall) for the T-stage classifier are shown in Table

3.

Training	Precision	Recall	F ₁ score
T1a	0,50	0,50	0,50
T1b	0,92	0,89	0,91
T1c	0,90	0,86	0,88
T2	0,67	1,00	0,80
T2a	0,85	0,91	0,88
T2b	1,00	0,89	0,94
T3	0,82	0,85	0,84
T4	0,83	0,83	0,83

Validation	Precision	Recall	F ₁ score
T1a	1,0	0,50	0,67
T1b	0,83	0,65	0,73
T1c	0,82	0,88	0,85
T2	0,60	1,00	0,75
T2a	0,83	0,89	0,86
T2b	0,88	0,91	0,89
T3	0,90	0,88	0,89
T4	0,84	0,89	0,87

Table 3. Precision, recall and F₁-scores T-substage for the training set and validation set.

In Table 4, errors made in the training set and validation set have been grouped into specific categories. In total, 27 (13.5%) errors were made in the training set and 35 (15.6%) errors in the validation set. Most errors were scored in the ambiguity category for the training (48.1%) as well as the validation set (51.4%).

Error Group	Error Type	Description	Training (n = 200)	Validation (n = 225)
Data selection	Sectionizer	Detects information in wrong subheadings	1	3
	Missing blacklist synonyms	Falsely matched / falsely not excluded	0	5
Context	Context missing	Context not matched because of missing modifier	1	0
	Context mismatch	Context mismatch, wrong modifier detected	1	3
Concept Matching	Measurement Extractor	e.g. using abbreviations (e.g. (AP) x (TVR) x (SI))	1	2
	Complexity	T ₄ multiple lobes	2	1
	Ambiguity	Confusion between node and mass (specific site: hilar)	4	7
		Non-specific		4

	Missing concepts synonyms	Lobulated	1	0
		Cystic	2	0
		Pleural thickening	1	0
		Spinal metastasis	1	0
		Costal involvement	0	1
		Supraclavicular extension	0	1
Reporter	Wrong input	Different sizes for the same tumor, no unit (mm/cm) present, size for tumor and atelectasis	7	2
		Satellite node	1	1
	Total Errors		27	35

Table 4. T-stage errors by category for training and validation sets.

Graphical User Interface

By using this tool (Fig. 2), the report and the tumor specific concepts are shown in a structured layout. Items that are present, missing or incorrectly stated can now be visualized. For instance, when tumor size or its unit is missing, size is not mentioned in the user interface and final T-stage will not be extracted properly. To increase its functionality, providing feedback and correcting errors, it is possible to adjust the proposed T-stage by changing the concepts found using drop down menus overruling the algorithm. In addition, this adjusted report can be saved anonymously and can be used as feedback to further improve the algorithms' future accuracy. As such, this tool can also function as a corpus builder when reports are being created. A consequence of the language-independent output format of our algorithm is that the Dutch algorithm is also available in the same GUI. The language can be set by clicking on a button.

Discussion

This study was performed to transfer and externally validate the Dutch rule-based pulmonary tumor T-stage NLP algorithm in an English cohort with the use of a GUI. Accuracy scores in this English study were similar to the scores found in the Dutch

cohort. The results confirm that the used strategy according to *size*, *involvement* and *presence* is viable and can also be implemented in a different language other than Dutch. The approach to find appropriate synonyms according to the Dutch outcomes (i.e. synonyms and found SNOMED-CT terms) was sufficient to get started. Adjusting the synonyms, without changing the algorithm itself, was enough to increase its accuracy. This again shows that the rule-based approach is very promising and can be implemented with a fairly high accuracy. Especially when taking into account that collecting data, and training and validation of the algorithm was done in roughly four weeks.

When looking at the separate F₁ scores, outcomes are slightly higher in the training set, but still have overall decent scores. The confusion matrices show that this algorithm tends to slightly overstage lower T-stages (up to T2a) and slightly understage the higher T-stages (from T2b onwards). This can be partly explained by the fact that it is more plausible to overstage a lower T-stage and understage a higher T-stage. However, as described in the following sections, this is most likely the result of difficulties experienced with the *overall reporting differences* and can be further explained with highlighting the *errors made by category* and *improving the algorithm* as stated below.

Overall reporting differences

One of the most important things was to find differences in reporting manner between the Dutch and English setting. Therefore, it was necessary to analyze the reports on a fundamental basis to find differences in reporting manner and used vocabulary in addition to the local used subheadings and layout. Because this cohort also used PET-CT scans in addition to CT scans, subheadings had to be added and the processing format had to be adjusted.

When looking at the reporting manner and the vocabulary used, the description of lymph node locations was found to be different in English as they are described in words (e.g. subcarinal) and not by numbers (e.g. level 7) as commonly done in Dutch reports. Another important finding was that the word ‘involvement’ and its conjugations (‘involv’-ing) was not exactly interchangeable, because involving has a more ambiguous meaning in English than the Dutch word for involvement (‘ingroeï’ - extension), which is very specific.

Furthermore, Dutch reports mention involvement when involvement is certain or suspected with a high level of certainty. Possible but less certain involvement is commonly not mentioned. The included English reports use more frequently terms to describe possible invasion without stating the exact certainty of the invasion with words such as ‘extending towards’, ‘abutting’ or ‘in close relation with the tumor’. As the algorithm was trained on matching the specific concept for invasion and the invaded concept, outcomes were more often falsely matched, in turn leading to false positive results.

Errors made by category

Data selection

In this category especially the blacklist in the validation set was not sufficient enough, which resulted in five entities that were falsely classified as tumor, but were benign lesions (for instance a benign kidney cyst with a size). As a consequence, these benign entities and its sizes were falsely seen as a tumor, resulting in an overestimation of the actual tumor size.

Context matching

Errors made in this category were due to a mismatch between the concept and the context. This happened for instance when a report lacked tumor dimensions, but instead was called ‘large’. Because tumor size is needed for the algorithm to appreciate something as a mass, this mass was missed. Another difficulty was to find sufficient synonyms for the ambiguous term atelectasis, which can be referring to either a post-obstructive atelectasis, which is a concept for the item *presence* in T2 tumors, or a regular seen non-tumoral gravity related atelectasis. Specific atelectasis related adjectives (basal, bilateral, subsegmental, etc.) were used to exclude gravity related atelectasis, as was done in the Dutch approach. However, this could not prevent some mismatches and overstaging of T1 tumors.

Concept matching

Most errors are related to this category with several subcategories, of which ambiguity is the largest contributor. Errors in this category were mainly due to difficulties in differentiating a lymph node from a tumor mass and difficulties in finding its proper location. This is especially true for the hilar region. For instance, a lymph node can be described as 'a (lymph nodal) subcarinal mass' and a tumor as 'a (peri)hilar node', making its exact location, and whether this is a primary tumor, less clear. Inserting more specific synonyms for involvement (i.e. involvement in(to)) and specific terms for lymph node and mass location (i.e. subcarinal lymph node) increased accuracy, but could not solve this problem entirely. This was in Dutch reports a lesser problem because lymph node levels are mostly mentioned by level number.

The error type missing/misuse concept synonyms is of particular interest because it shows difficulties caused by the rule-based algorithm approach best. One error in this subcategory was made because there was a size at an involvement concept (visceral pleura) that therefore could not be blacklisted (e.g. 'pleural thickening of 8,6 cm'). Also the opposite errors existed in cases, in which a cystic pulmonary tumor was missed because the word cyst was blacklisted to not falsely match a renal cyst. In addition, it was not possible to differentiate osseous destruction caused by the primary tumor from destruction caused by a vertebral metastasis. The complexity subcategory and errors made by the measurement extractor have similar difficulties in which it is difficult to match a different tumor in the same lobe of the ipsilateral lung or match a tumor size when the size is written in an uncommon format.

Reporter

This category included errors which can be explained by stating wrong tumor size, or mentioning it twice, or incorrectly reporting the presence of a satellite nodule. As the algorithm demands a unit for every size with its size is correlated to the stated lobe, it is possible that a tumor without a unit is missed, a tumor with two different sizes is overestimated and a satellite nodule in a different lobe is missed. As such, correctly stated input is of great importance.

Improving the algorithm

The described rule-based algorithm is promising, but this approach is a tradeoff between missing a lung tumor and finding a false mass. The rule-based nature of the use of the sectionizer and regular expressions is not extensive enough to exclude the ambiguities, nor to find sections when those were not present in the training set. Furthermore, the context analysis does not search for dependency relations resulting in mismatches between concept and context. In addition, the rule-based approach does not seem extensive enough, as the T-stage based on only tumor size has an accuracy of respectively 0.78 and 0.76 in the training and validation sets. The additional set of rules improves the outcome only by 0.08-0.09.

Although NLP can be successfully applied in free text reports, its accuracy will benefit from increasing levels of structure and standardization in the report. In addition, machine learning is thought to increase the accuracy score by finding more related synonyms based on a larger amount of data. This can be achieved by using, for instance, word embeddings. This allows for more extensive analysis of the context, because specific concepts are often embedded by the same set of modifiers.

Although machine learning may be a promising addition, it requires much more annotated data for training purposes. Availability of these large amounts of specific data is sometimes an issue, especially at the beginning of a new measurement method or a new edition of the TNM-staging. In addition, extracting and labeling large amounts of data is expensive and time consuming. Therefore, it is important to learn from this baseline study and explore where exactly implementation of machine learning or deep learning methods could increase outcomes. Focusing on finding accurate synonyms (e.g. gravity dependent atelectasis/non-oncological atelectasis), distinguishing tumor from lymph node and accurate matching of contextual information to the right concepts might be a way to improve the algorithm. This hybrid approach could increase outcomes more efficiently, without the need to annotate a vast amount of data. This could result in lower costs and speeds up the availability of these algorithms. In addition, less *specific* data can be used to train the algorithm because only the experienced difficulties need to be trained. For instance, non-oncological atelectasis is also mentioned in non-oncological CT or PET-CT scans.

Clinical significance and future perspectives

Future work should focus on improving the algorithm, but research can also be aimed at how such algorithms can help with restaging tumor classifications across staging editions or how a classification GUI can be implemented in clinical practice.

In this study, the GUI is only used for finding, analyzing and adjusting errors during training. However, this tool can also be implemented for (live) staging during the reporting process. When connected to the directory in which reports are made, (live) staging aids the reporter in increasing accuracy, completeness and quality of the report by making sure that specific concepts are mentioned in the free text report by looking at the (already filled in) structured format. The GUI can notify the reporter with a pop-up screen that pivotal information is missing. In this study, 8 and 3 reporter related errors were found in, respectively, the training and validation set. These could be prevented when information was checked before finishing the report. The use of this algorithm with the GUI could have increased the report accuracy (i.e. quality of the report) by 1.5-4.0%. As such, the GUI might lead to better reports and perhaps also nudge the radiologist to more structured and standardized reporting as they see the direct effect of that in the GUI.

Moreover, the potential of these types of algorithms will be further enhanced when they are used in less difficult settings, such as automatic extraction of the TIRADS (Thyroid Imaging Reporting And Database System) classification of thyroid nodules as described in thyroid ultrasounds. Automatically stating the Bosniak classification on CT scans used to describe cystic renal masses may be another example. When we can also combine Artificial Intelligence (AI)-based automated image extraction information tools (e.g. tumor size extractor), it might be possible to prefill the radiological report and assist the reporter and the algorithm further.

A different opportunity of NLP is to extract certain endpoints, such as the presence of a specific disease or important or incidental findings. This can be used to (semi)automatically warn the referring specialist or plan a follow-up appointment.

As such, applying these algorithms in clinical practice can be complementary to structured reporting in radiology. It automatically checks the free text report for specific items and converts these items into a structured format, without extensively changing or interfering the way of reporting. This is especially of importance in times of

datafication and increased need for data standardization as promoted by the ESR and RSNA. It shows that also NLP or rule-based algorithms can reinforce the radiologist and their reports, further supporting the reporting process.

Limitations

A limitation of using a rule-based approach building this T-stage algorithm is that specific boundaries had to be determined if those were not specified by the TNM. For instance, it was necessary to specify the size of the node in the ipsilateral side of the main tumor in a different lobe for T₄ stage (>1 cm) and the size for a different tumor in a different lobe (>1 cm).

Another limitation was that we had to determine the strictness of the algorithm, and more specifically on concepts such as involvement or presence. It is debatable whether only obvious invasion should be accounted for an involved concept or whether terms like 'likely' or 'probably' should be added to the invaded concepts. However, the presented rule-based algorithm can be configured. Furthermore, the obtained T-stage by this algorithm is a radiological T-stage. This may be different from the final T-stage, which generally also requires additional clinical information.

Lastly, the T-stage scoring process was done by one author (JMN). In case of uncertainty and/or ambiguity, a second author (JW) was consulted, after which consensus was reached between two authors. Although future validation studies should also look at aspects of interrater variability, the primary goal of the current study is to explore whether the Dutch algorithm could be useful when translated into English.

Conclusion

NLP is a promising tool that can be used in extracting specific information from radiological reports concerning T-stage in pulmonary oncology. The used Dutch algorithm could be successfully translated and validated in an English dataset and this will likely be feasible for other languages as well. Focused implementation of more machine learning strategies and the use of a Graphical User Interface should lead to higher accuracy, as an effect of better report quality.

References

1. Pinto dos Santos D. The Value of Structured Reporting for AI. In: Ranschaert ER, Morozov S, Algra PR, editors. *Artificial Intelligence in Medical Imaging: Opportunities, Applications and Risks*. Cham (Switzerland): Springer Nature; 2019.
2. Kreimeyer K, Foster M, Pandey A, Arya N, Halford G, Jones SF, et al. Natural language processing systems for capturing and standardizing unstructured clinical information: A systematic review. *J Biomed Inform*. 2017;73:14-29. doi: 10.1016/j.jbi.2017.07.012.
3. Kehl KL, Elmarakeby H, Nishino M, Van Allen EM, Lepisto EM, Hassett MJ, et al. Assessment of Deep Natural Language Processing in Ascertaining Oncologic Outcomes From Radiology Reports. *JAMA Oncol*. 2019;5(10):1421-1429. doi: 10.1001/jamaoncol.2019.1800.
4. Cheng LT, Zheng J, Savova GK, Erickson BJ. Discerning tumor status from unstructured MRI reports: completeness of information in existing reports and utility of automated natural language processing. *J Digit Imaging* 2010;23(2):119-132. doi: 10.1007/s10278-009-9215-7.
5. Jain NL, Friedman C. Identification of findings suspicious for breast cancer based on natural language processing of mammogram reports. *Proc AMIA Annu Fall Symp*. 1997:829-833.
6. Mamlin BW, Heinze DT, McDonald CJ. Automated extraction and normalization of findings from cancer related free-text radiology reports. *AMIA Annu Symp Proc*. 2003;2003:420-424.
7. Brierley J, Gospodarowicz MK, Wittekind C, editors. *TNM classification of malignant tumours*. 8th ed. Chichester: John Wiley & Sons Inc; 2017.
8. Nobel JM, Puts S, Bakers FCH, Robben SGF, Dekker ALAJ. Natural Language Processing in Dutch Free Text Radiology Reports: Challenges in a Small Language Area Staging Pulmonary Oncology. *J Digit Imaging*. 2020;33(4):1002-1008. doi: 10.1007/s10278-020-00327-z.
9. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J Clin*. 2021;71(3):209-249. doi: 10.3322/caac.21660.
10. Hassanpour S, Langlotz CP. Information extraction from multi-institutional radiology reports. *Artif Intell Med*. 2016;66:29-39. doi: 10.1016/j.artmed.2015.09.007.
11. Kohli M, Alkasab T, Wang K, Heilbrun ME, Flanders AE, Dreyer K, et al. Bending the Artificial Intelligence Curve for Radiology: Informatics Tools From ACR and RSNA. *J Am Coll Radiol*. 2019;16(10):1464-1470. doi: 10.1016/j.jacr.2019.06.009.
12. European Society of Radiology (ESR). ESR paper on structured reporting in radiology. *Insights Imaging*. 2018;9(1):1-7. doi: 10.1007/s13244-017-0588-8.
13. European Society of Radiology (ESR). ESR communication guidelines for radiologists. *Insights Imaging*. 2013;4(2):143-146. doi: 10.1007/s13244-013-0218-z.
14. European Society of Radiology (ESR). Good practice for radiological reporting. Guidelines from the European Society of Radiology (ESR). *Insights Imaging*. 2011;2(2):93-96. doi: 10.1007/s13244-011-0066-7.

15. Faggioni L, Coppola F, Ferrari R, Neri E, Regge D. Usage of structured reporting in radiological practice: results from an Italian online survey. *Eur Radiol.* 2017;27(5):1934–1943. doi: 10.1007/s00330-016-4553-6.
16. Powell DK, Silberzweig JE. State of structured reporting in radiology, a survey. *Acad Radiol.* 2015;22(2):226–233. doi: 10.1016/j.acra.2014.08.014.
17. Weber TF, Spurny M, Hasse FC, Sedlaczek O, Haag GM, Springfield C, et al. Improving radiologic communication in oncology: a single-centre experience with structured reporting for cancer patients. *Insights Imaging.* 2020;11(1):106. doi: 10.1186/s13244-020-00907-1.
18. Marcovici PA, Taylor GA. Journal Club: Structured radiology reports are more complete and more effective than unstructured reports. *AJR Am J Roentgenol.* 2014;203(6):1265–71. doi: 10.2214/AJR.14.12636.
19. Côté RA, Robboy S. Progress in medical information management. Systematized nomenclature of medicine (SNOMED). *JAMA.* 1980;243(8):756–762. doi: 10.1001/jama.1980.03300340032015.
20. Honnibal M, Montani I. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. To appear: 2017.
21. Chapman BE, Lee S, Kang HP, Chapman WW. Document-level classification of CT pulmonary angiography reports based on an extension of the ConText algorithm. *J Biomed Inform.* 2011;44(5):728–37. doi: 10.1016/j.jbi.2011.03.011.
22. Puts S, Nobel JM. Medical narrative to structure: maastroclinic/medstruct. maastroclinic; 2019.

Supplementary material

Appendix 1: Annotation guidelines

- Stated as being certain
- Secondary tumor ipsilateral: size > 1,0 cm
- Atelectasis by tumor
- Satellite nodules only when in the same lobe

Appendix 2: Concept synonyms

Regular expressions used for classification concepts tumor and involvement, corresponding SNOMED-CT concepts.

General

concept	regular expression	SNOMED CT concept
tumor	'(tumor tumour carcino malign)'	108369006 Neoplasm (morphologic abnormality)
involvement	'affecting attacking injuring destruct ingrowth growth (extension extends extending expansion)[]*(in into) (involvement infiltration)[]*of involve invading in invas invades'	248448006 Involved (qualifier value)
lymph nodes	'lymph lymph.*node'	59441001 Structure of lymph node (body structure)

T2 Presence

concept	regular expression	SNOMED CT concept
main_bronchus	'(central main first).*bronch'	102297006 Main bronchus structure (body structure)
visceral_pleura	'pleura'	81623005 Visceral pleura structure (body structure)

T2 Involvement

concept	regular expression	SNOMED CT concept
atelectasis	'atelect collapse'	46621007 Atelectasis (disorder)
obstructive_pneumonitis	'obstructive pneumoni infect.*chang obstructive pneumonitis obstructive infectious disease obstructive pneumonia'	205237003 Pneumonitis (disorder)

T3 Involvement

concept	regular expression	SNOMED CT concept
chest_wall	'chest.*wall thorax.*wall rib costa'	78904004 Chest wall structure (body structure)
nervus_phrenicus	'nervus.*(phrenicus frenicus) phrenic nerve nervus phrenicus'	50230006 Structure of phrenic nerve (body structure)
parietale_pericard	'pericard'	76848001 Pericardial structure (body structure)

T3 Presence

concept	regular expression	SNOMED CT concept
satellite_nodule	'satellite nodule satellite nod satellite lesion'	396408009 Specimen involvement by satellite nodule(s) present (finding)

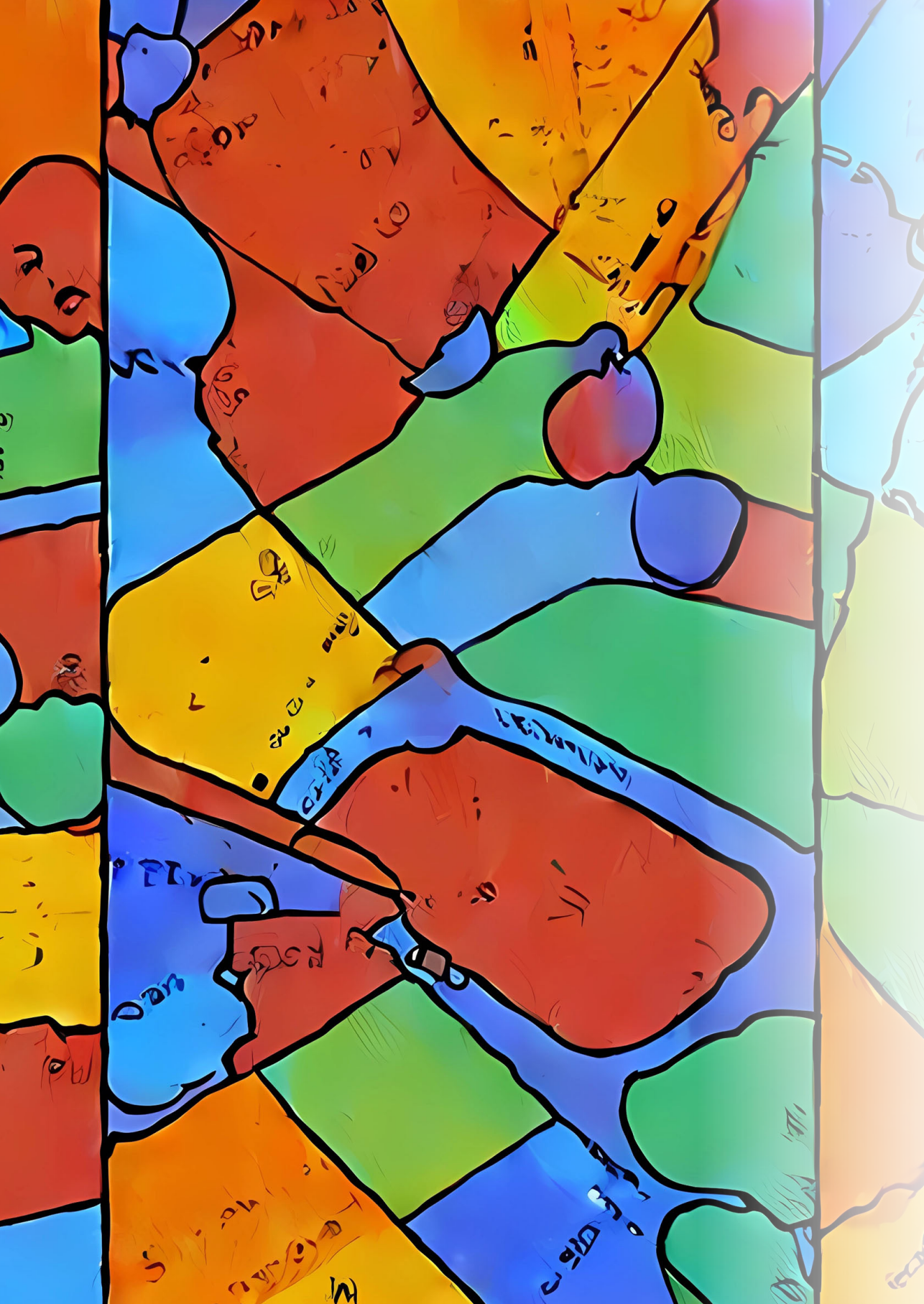
T4 Involvement

concept	regular expression	SNOMED CT concept
diaphragm	'diaphragm'	5798000 Diaphragm structure (body structure)
mediastinum	'mediast mediastinum mediastinal fat'	72410000 Mediastinal structure (body structure)
heart	'heart cor cardial'	80891009 Heart structure (body structure)
great vessels	'great vessel great vessels central vessels central vessel aorta vena cava VCS main pulmonary artery main pulmonary vein'	3711007 Structure of great blood vessel (organ) (body structure)
trachea	'windpipe \btrachea\b \btracheal\b'	44567001 Tracheal structure (body structure)

recurrent_laryngeal_nerve	'recurrent laryngeal nerve laryngeal nerve nervus laryngeus recurrens'	731050007 Entire recurrent laryngeal nerve (body structure)
oesophagus	'oesophagus esophagus'	32849002 Esophageal structure (body structure)
vertebral body	'vertebral body vertebra spine spinal'	3572006 Structure of body of vertebra (body structure)
carina	'\bcarina\b \bcarinal\b'	28700002 Structure of carina of trachea (body structure)

T4 Tumor in different lobes

concept	regular expression	SNOMED CT concept
superior_lobe_left	'superior left lobe superior lobe left upper left lobe left upper lobe upper left lobe apical left lung apical lung left upper lobe of left \bLUL\b \bSLL\b'	44714003 Structure of upper lobe of left lung (body structure)
middle_lobe	'middle lobe center lobe lobus intermedius \bML\b \bRML\b'	72481006 Structure of middle lobe of right lung (body structure)
inferior_lobe_right	'inferior right lobe inferior lobe right lower right lobe right lower lobe lower lobe right basal lung right basal right lung \bRLL\b \bILR\b'	266005 Structure of lower lobe of right lung (body structure)
inferior_lobe_left	'inferior left lobe inferior lobe left lower left lobe left lower lobe lower lobe left basal lung left basal left lobe \bLLL\b \bILL\b'	41224006 Structure of lower lobe of left lung (body structure)



Chapter 6:

Automated pulmonary oncology staging from free text radiological reports: extending the Dutch algorithm towards full utilization

Sander Puts, J. Martijn Nobel*, Catharina M.L. Zegers,
Iñigo Bermejo, Simon G.F. Robben, André L.A.J. Dekker*

**Equally contributing authors*

Adapted from JMIR Formative Research (JFR) (2023)

Abstract

Natural Language Processing (NLP) is thought to be a promising solution to extract and store concepts from free text in a structured manner for data mining purposes. This is also true for radiology reports, which still consist mostly out of free text. Accurate and complete reports are very important for clinical decision support, for instance in oncological staging. As such, NLP can be a tool to structure the content of the radiology report, thereby increasing the report's value.

This study describes the implementation and validation of an N-stage classifier for pulmonary oncology. It is based on free text radiological chest Computed Tomography (CT) reports according to the Tumor Node Metastasis (TNM) classification, which has been added to the already existing T-stage classifier to create a combined TN-stage classifier.

SpaCy, PyContextNLP and Regular Expressions (RegEx) were used for proper information extraction, after additional rules were set to accurately extract N-stage. The overall TN-stage classifier accuracy scores were 0.84 and 0.85 for, respectively, the training ($n = 95$) and validation ($n = 97$) sets. This is comparable to outcomes of the T-stage classifier (0.84-0.88).

This study shows that a rule-based approach is feasible, but heterogeneity of the radiological reports makes it difficult to improve the outcomes more. Machine learning is expected to be capable of increasing accuracy further, in which a hybrid approach could be the preferred way to get the best results.

Introduction

Staging oncological patients is of utmost importance to determine the most appropriate treatment regime to ensure the best outcome for the patient. The Tumor Node Metastasis (TNM) classification system is internationally accepted as a standard for proper staging of cancer patients [1]. Radiological imaging by means of a chest Computed Tomography (CT) scan is an important pillar for the TNM classification in clinical practice. Because the radiological report is the way to communicate observations to referring clinicians, the content of the report needs to be complete and

accurate [2, 3, 4]. Specifically, the Tumor (T), the Node (N) and the Metastasis (M) status should be known. However, the radiological report is in most cases still a free text report in which layout, structure, readability and accuracy largely depends on the reporter.

Natural Language Processing (NLP) can be applied to extract specific information from free text. This can also be applied to radiological reports when, for instance, specific coding and structured reporting are not used [5, 6]. Already several studies have been performed using NLP in radiology and implementation in clinical practice seems just a matter of time [7, 8, 9]. Examples of specific oncological NLP implementations on radiology reports or the Electronic Health Record (EHR) are in oncological follow-up, tumor recurrence rates, follow-up of a critical oncological finding and uses for cancer registries [10, 11, 12, 13]. Also non-oncological studies have been performed using NLP to search for specific statements from pulmonary angiography reports, imaging reports of subdural hematoma in the acute setting or, more generally, to extract recommendations from radiology reports [14, 15, 16].

NLP has also been used in a recent and ongoing transnational project to extract the stage in pulmonary oncology from free text radiological chest CT scan reports [17]. The overall goal is to build a language independent algorithm that can extract pulmonary oncology staging according to the TNM classification. In prior work, a rule-based NLP algorithm was trained and validated on Dutch radiological reports before it was translated and validated on English reports, which showed an accuracy rate for T-stage ranging between 0.84-0.88. This rule-based approach is thought to be the easiest way to accurately determine the oncologic stage, as TNM is already a rule-based system. When for instance only machine learning (ML) strategies for staging were used, apart from the issue of correctly finding the specific concepts, the algorithm also needs to extract the set of rules of each concept from the training data, which requires a very large amount of data.

For adequate staging, the N-stage should also be known. We hypothesize that, as the items to build the N-stage should be mentioned in the same radiological staging report as used to classify the T-stage, it should be possible to accurately extract the N-stage

from the report using a similar process as previously used for the T-stage. This paper describes the process of training and validation of extraction of the N-stage of pulmonary oncology of Dutch free text radiological chest CT reports and discusses whether this is a feasible tool in addition to the already validated rule-based T-stage algorithm.

Methods

Corpus description

Ethical approval was waived at the participating institute. For this study, radiological reports of diagnostic chest CT scans used for the staging of pulmonary oncology were used. The training and validation sets consisted of respectively 95 and 97 reports. Reports were included when a primary pulmonary malignancy was described by a radiologist. The included free text radiological reports have been constructed by several different radiologists, other than the authors, using a speech recognition tool (G2 Speech). Exclusion criteria were 1) restaging and follow-up reports, 2) cases with two primary tumors and 3) incomplete reports. The included reports were independently classified by two authors (JM, SP) according to the 8th TNM classification system [1]. For every report, the T-stage and N-stage was labeled. Because TNM stage was not specified in the radiological report, this had to be done manually. Annotation guidelines were set for proper and consistent labeling (see Appendix 1: Annotation guidelines). Tumor stage characteristics of both groups are shown in Table 1 (Cohort composition of the training and validation set). The layout of the included reports differed and contained one or more of the following subheadings: clinical details, description of the modality, report, body part and impression.

The training set was used to identify the content of the radiological report to find the appropriate synonyms used for reporting N-stage. These synonyms were used to build new N-staging rules which were incorporated in the existing T-stage rule-based algorithm.

	Training (n = 95)	Validation (n = 97)
T1aNo	0	0
T1aN1	0	0
T1aN2	1	0
T1aN3	0	0
T1bNo	7	2
T1bN1	2	0
T1bN2	1	1
T1bN3	0	1
T1cNo	7	9
T1cN1	0	3
T1cN2	2	5
T1cN3	3	1
T2No	0	1
T2N1	0	0
T2N2	4	3
T2N3	1	1
T2aN0	4	5
T2aN1	2	2
T2aN2	2	3
T2aN3	3	1
T2bNo	4	2
T2bN1	0	2
T2bN2	4	6
T2bN3	3	2
T3No	5	5
T3N1	1	0
T3N2	6	9
T3N3	4	6
T4No	7	8
T4N1	0	2
T4N2	13	11
T4N3	9	6

Table 1. Cohort composition of the training and validation sets.

Determining T-stage

In this study, the same rule-based TNM T-stage algorithm was used as published earlier, with pre-processing steps, such as sectionizing, text cleaning, extraction of numbers and accurate sentence splitting [17]. The processing steps were based on the extraction of

three items important for T-staging using Regular Expressions (RegEx): *size*, *presence* and *involvement* (Fig. 1). Outcomes were used for T-staging the tumor (Appendix 2).

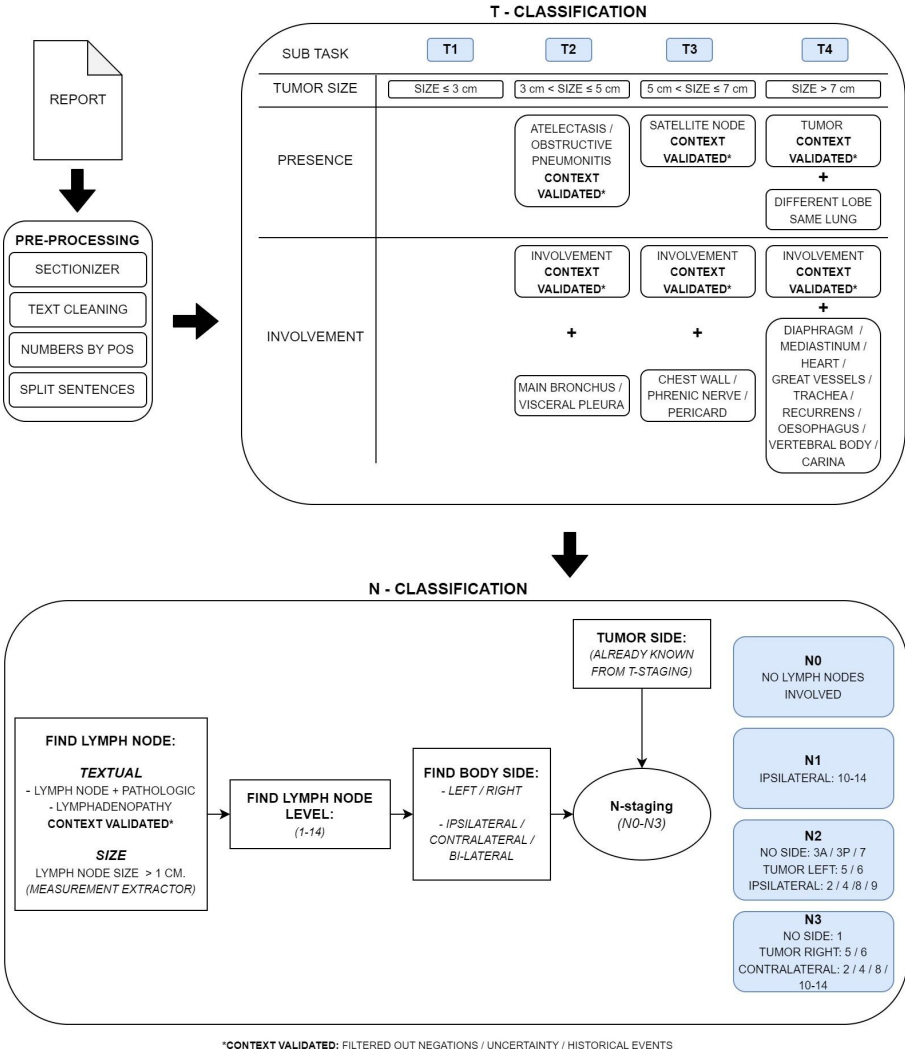


Figure 1. Schematic overview rule-based T and N-staging algorithm. In the pre-processing step, raw data of the report is prepared for actual processing. The processing is divided in T-stage and N-stage in which several subtasks are displayed to finally stage the pulmonary tumor and pulmonary lymph nodes.

Determining N-stage

For extracting the N-stage of pulmonary oncological cases the 8th TNM classification was analyzed in detail and four items were recognized as important: (*pathological*) *lymph node*, *lymph node level*, *lymph node side* and *tumor side* (Fig. 1). To accurately stage the N-stage, the described lymph nodes had to be found and matched with its potential context first, to know whether or not the lymph node was a pathological lymph node. Therefore, synonyms of N-specific concepts, such as ‘lymph nodes’ and ‘pathological’, had to be found to build a specific RegEx per concept. Therefore, lymph node specific rules had to be built.

For N-staging it was necessary to look more extensively at the relation between ‘*context target*’ and the ‘*context modifiers*’, because it appeared that a pathological lymph node was less specifically mentioned in the report than the primary tumor. A target could, for instance, be the concept for the word “lymph node” and the modifier the adjective, stating it is “enlarged”. Furthermore, an enlarged lymph node could be described by only text, but can also be highlighted by quantifying its enlarged size. This resulted in finding three distinct ways of mentioning the pathological lymph node in which, 1) “lymph node” and “pathological lymph node” had to be extracted, 2) “lymph node” and its pathological size and 3) “lymph node” and the word “pathological” had to be matched, and a specific RegEx had to be built per item. Regular expressions related to context, such as negations and uncertainty, could be reused from the T-staging process, but the additional category for “pathological” had to be added.

Subsequently, the lymph node level had to be found and, since there are fourteen different thoracic levels, a RegEx was built per level. Furthermore, the side of the tumor and the pathological lymph node had to be extracted to define the lymphadenopathy to be ipsilateral, contralateral or bilateral. This was not necessary for extraction of the T-stage, and specific rules for sentence analysis had to be set. Furthermore, the size of the lymph node was extracted by the measurement extractor component, which uses the number category of the open-source part-of-speech (POS) tagger as input. Finally, the tumor side was matched to the side of the pathological or enlarged lymph node and used for definitive N-staging (Appendix 2).

Statistical analysis

For both the training and the validation set, the substage accuracy scores were calculated separately for the T-stage and the N-stage. Furthermore, the combined accuracy score (TN-stage) was scored for the training and validation sets. To find out whether the N-stage extension in this TN-classifier compromised the T-stage outcomes, also the earlier version of the algorithm, which was a T-stage classifier only, was run on the training and validation sets. In addition, the accuracy score was calculated when only tumor size was taken into account.

Furthermore, the confusion matrices were built for the training and the validation sets to highlight the correlation between the *actual* N-stage and the *predicted* N-stage as well as for the *actual* TN-stage and the *predicted* TN-stage. In addition, the precision (i.e. specificity), recall (i.e. sensitivity) and F_1 measure (i.e. combined metric for precision and recall) for the combined TN-stage classifier was calculated for the training and validation sets. Different types of errors were grouped by category for further analysis: data selection, context extraction, concept extraction and reporter errors.

Graphical User Interface (GUI)

A GUI was built to highlight the TN-stage of the report in the staging screen (Fig. 2). When the N-rules were set, these rules have been implemented in this tool to help with the staging check by visualizing the scored TN-stage by the algorithm and compare those with the manually extracted TN-stage.

MEDSTRUCT-NLP

Assistance, Classification and Information Extraction for Medical Free-Text Reporting

View Annotated: Autocheck:

Annotated Report

Thoraxgrote massa zichtbaar in de linker bovenkwab met maximale diameter op 8 46 van 4.7 x 3.0 cm. **Mogelijke** ingroei in het mediastinum. **Satelliet laesies** zichtbaar op 8 41 met een grootte van 1,3 cm. **Lymfeklier** zichtbaar op station 7 met een lente van circa 5,2 cm. **Geen** lymfeklieren aan de contralaterale zijde zichtbaar. **Kleine** consolidatie in the middenkwab. **Geen** atelectase. Conclusietumor met satelliet laesies in de linker bovenkwab.

- Present
- Involved
- Context Modifier
- Context Target

Example text from Article-1 (NL) ▼

TNM-8 Lung

Nederlands Classify

T3 N2 M

Primary Tumor

4.7 cm Left Side

Present Involved

Satellite Nodule (T3) Ipsilateral Tumor (T4)

Lymph Nodes

7 Subcarinal x ▼

Figure 2. Graphical User Interface (GUI) MedStruct.

Results

The accuracy rates for the T-stage score were 0.87 and 0.92 for, respectively, the training ($n = 95$) and validation ($n = 97$) set. N-stage accuracy was, respectively, 0.96 and 0.92. The combined accuracy TN-stage scores were 0.84 and 0.85 for the training and validation sets (see Table 2: T, N and TN-stage classifier accuracy).

	<i>TN-classifier</i>		<i>T-classifier</i>	
	Training ($n = 95$)	Validation ($n = 97$)	Training ($n = 95$)	Validation ($n = 97$)
Accuracy T-stage (<i>T-substage</i>)	0.87	0.92	<i>N/A</i>	<i>N/A</i>
Accuracy N-stage	0.96	0.92	<i>N/A</i>	<i>N/A</i>
Accuracy TN-stage	0.84	0.85	<i>N/A</i>	<i>N/A</i>
Accuracy T-stage (<i>size only</i>)	0.80	0.81	0.76	0.79
Accuracy T-stage (<i>T-stage</i>)	0.89	0.93	0.82	0.86

Table 2. Accuracy scores of training and validation sets of the separate T-stage and N-stage and the combined TN-stage. For comparison, the T-classifier outcomes are shown for the current sets as well as the T-stage for only tumor size.

When looking at the earlier version of the algorithm, which only classified the T-stage, the accuracy score of this combined TN-classifier performed slightly better than the T-classifier. The confusion matrices are shown in Fig. 3, Fig. 4 and Fig. 5.

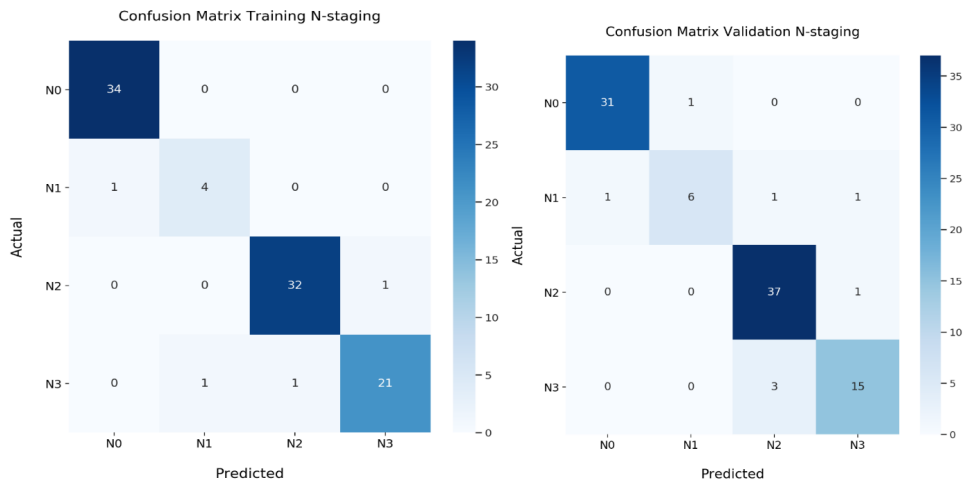


Figure 3. Confusion matrices of the N-stage classification only on the (a) training set and (b) validation set.

6

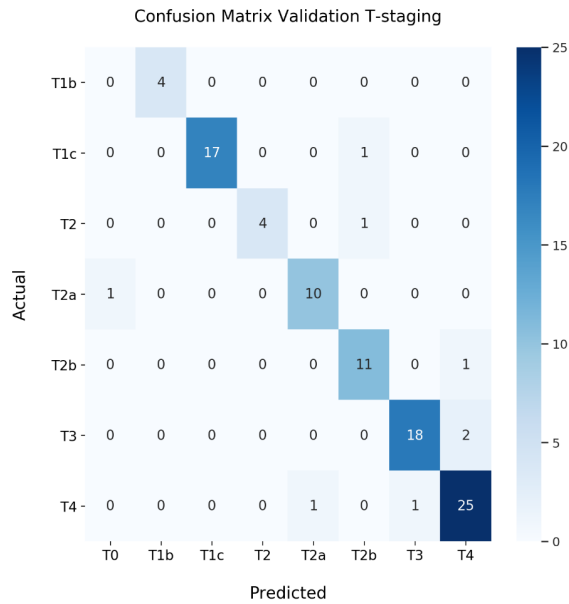
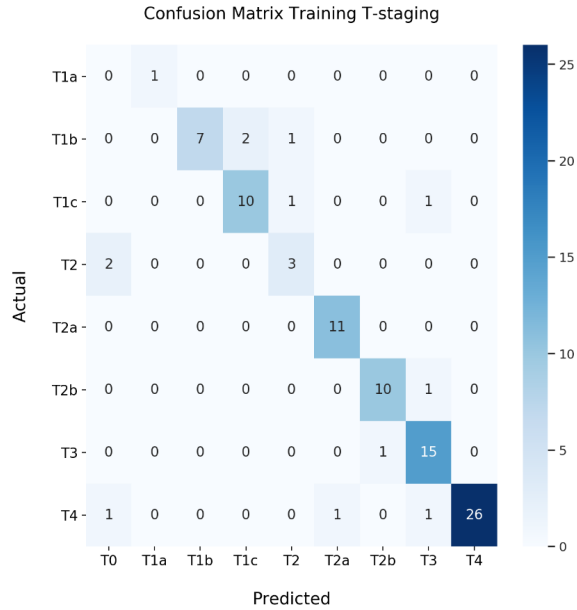
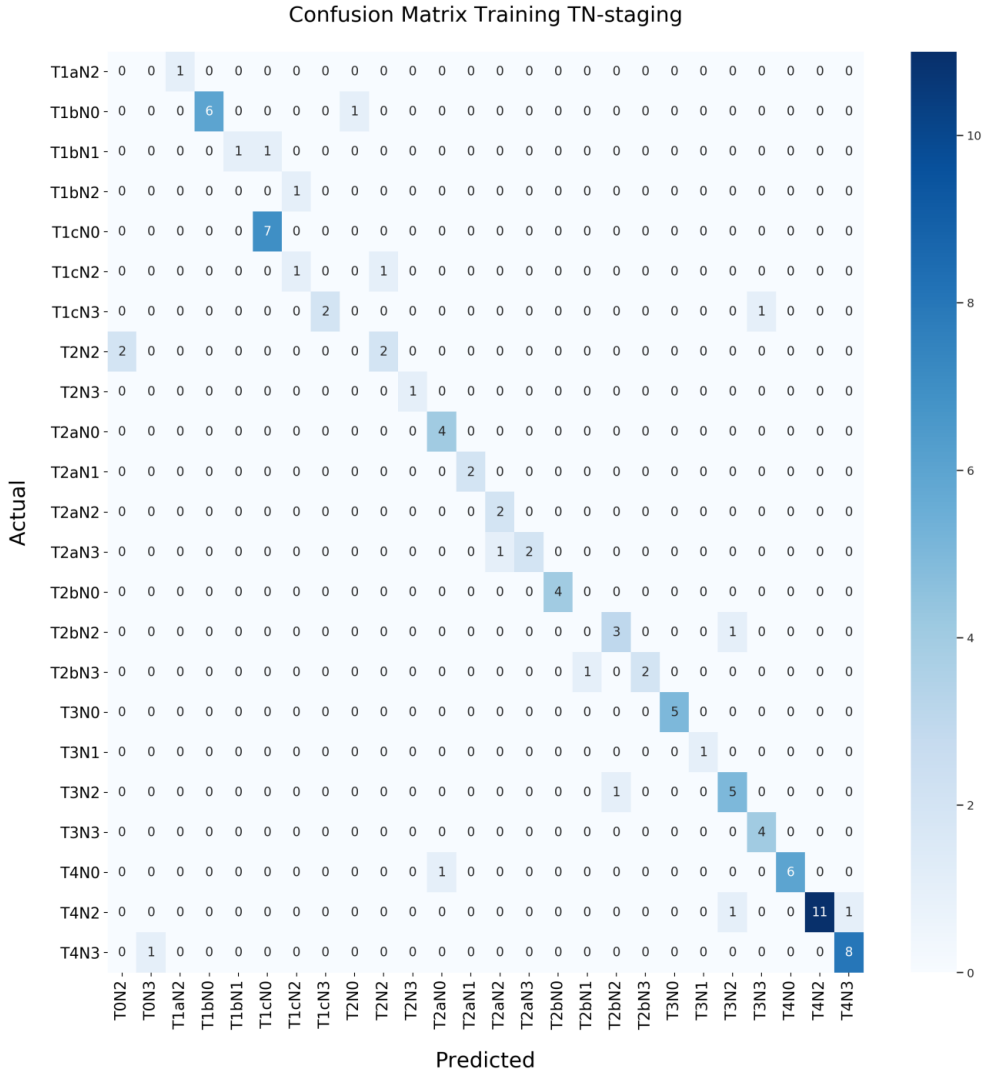
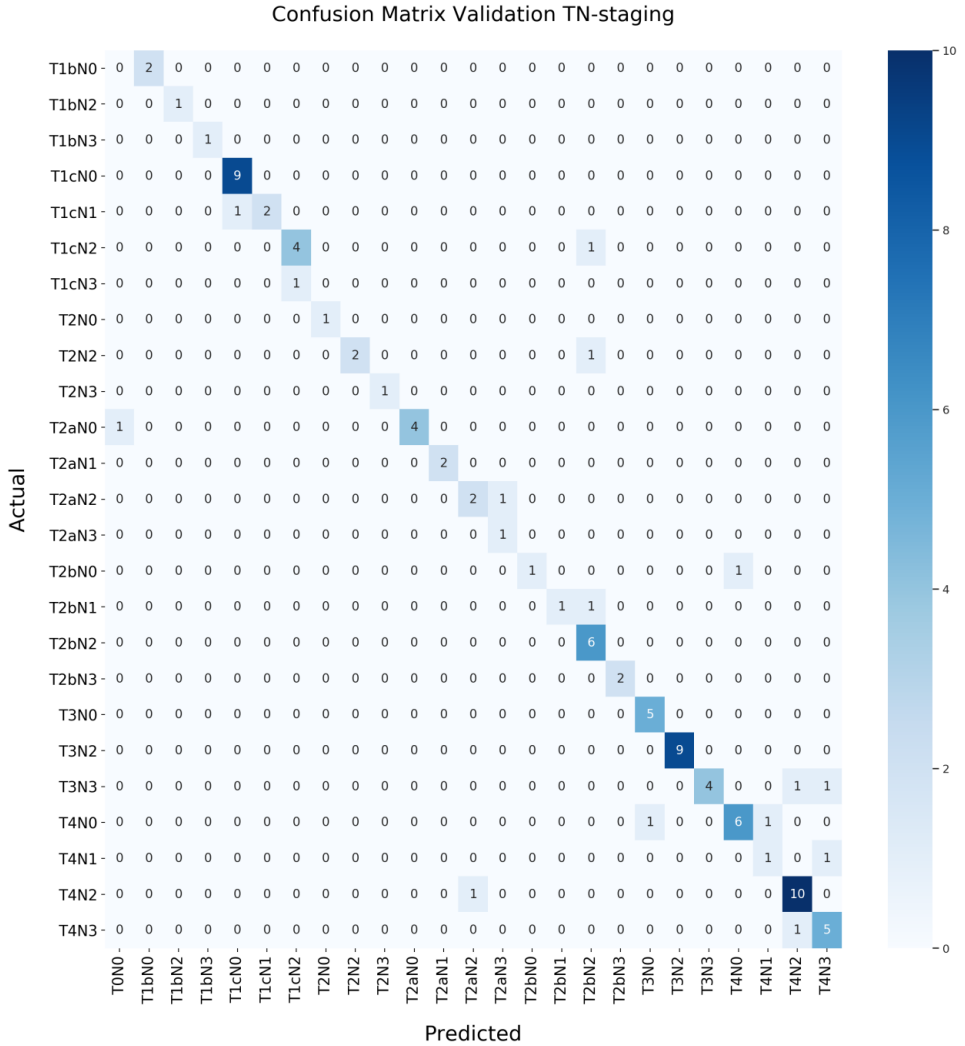


Figure 4. Confusion matrices of the T-stage classification only on the (a) training set and (b) validation set.



(a)

Figure 5a. Confusion matrices of the TN-stage classification on the (a) training set and (b) validation set.



(b)

Figure 5b. Confusion matrices of the TN-stage classification on the (a) training set and (b) validation set.

The precision, recall and F1 measure for the combined TN-stage classifier are shown in Table 3 (Weighted precision, recall and F1-scores of the TN-stage).

	Precision	Recall	F1 score
Training (overall)	0.89	0.84	0.86
Validation (overall)	0.87	0.85	0.84

Table 3. Weighted precision, recall and F1-scores of the TN-stage.

The errors found were categorized into specific subcategories as shown in Table 4. Sixteen errors were made in the training set and 16 in the validation set leading to 15 classification errors, with one error in both T and N-stage in one case, in both the training and the validation set.

Error Group	Error Type	Description	Training set (n = 95)	Validation set (n = 97)
Data selection	Sectionizer	Subheadings not present or falsely not found – falsely correlation tumor or nodal description	2 (1T, 1N)	2 (2T)
Context extraction	Missing	Context not matched because of missing / falsely matched modifier Stated uncertainty not found accurately	1 (1T)	3 (2T, 1N)
	Complexity	Context mismatch, wrong modifier detected: not or uncertainty	4 (3T, 1N)	
		Abdominal para-aortal lymph node		1 (1N)
Concept extraction	Missing	Synonym pathological		1 (1N)
	Ambiguity	Nodal description / station	2 (1T, 1N)	

		Tumor dependent atelectasis	2 (2T)	
		Pulmonary vein	1 (iT)	
	Complexity	Size description	1 (iT)	1 (iT)
		T4 multiple lobes – implicit mentioning	1 (iT)	1 (iT)
		Side implicit mention		1 (iN)
		Mention nodal status	1 (iN)	
Reporter	Wrong input	Typing / speech error	1 (iT)	2 (iT, iN)
		Incomplete node mentioning (location or pathological)		3 (3N)
		Inconsistent tumor location		1 (iT)
	Total errors		16*	16*

*16 errors in total, leading to 15 wrong classification scores

Table 4. TN-stage errors by category.

Discussion

This paper is another product of the effort to build a rule-based NLP algorithm to classify pulmonary oncology as reported in free text radiological CT chest staging reports according to the 8th TNM classification. In addition to the T-staging rules, specific N-staging rules were added to the algorithm in order to find four additional items necessary for proper N-staging: (*pathological*) *lymph node*, *lymph node level*, *lymph node side* and *tumor side*.

The accuracy scores for the N-stage were 0.96 and 0.92 in, respectively, the training and validation set. This goes to show that this rule-based approach and the rules set are viable for extraction of the items necessary for proper N-staging. From the combined TN-stage accuracy scores for the training and validation set, respectively 0.84 and 0.85, it can be observed that outcomes are a bit lower. However, taking into account that both the T-stage and the N-stage had to be correct, accuracy is still reasonably high, and comparable with outcomes of the T-stage alone (0.84-0.88). The outcomes of the accuracy score of the TN-classifier are comparable or slightly better than the accuracy

score of the T-classifier only, showing that addition of the N-stage rules did not interfere with the overall outcome.

Looking at the combined outcomes of the training and validation confusion matrices, it can be observed that the N-stage outcome was understaged in 7 cases and overstaged in 5. T-stage outcome was understaged in 9 cases and overstaged in 11 cases out of the grand total 192 cases. The TN-stage confusion matrices show that in both sets 15 cases were wrongly classified, and that in total 14 cases were understaged and 16 overstaged. Also, the errors made were equally divided between both sets. When looking at the error categories, a total of 20 errors were made in the T-stage classification and 12 in the N-stage classification (Table 4). In one case in both groups the N-stage and the T-stage were both falsely staged. Overall, many different errors occurred, which shows the heterogeneity of the reports, and hence the extent of the task to tackle and optimize this rule-based approach.

The difference between the high N-stage accuracy scores and the lower TN-stage accuracy scores are slightly compromised by difficulties still experienced by the T-staging rather than by N-staging difficulties (see Table 4: TN-stage errors by category). This can be explained by the fact that the T-staging process is more difficult to accurately perform with a rule-based approach and may therefore be less reliable than the N-staging. This is not surprising when looking at the number of substages used in the T-stage compared to the N-stage, with 8 substages for T-staging (T_{1a} – T₄) versus 4 substages for the N-stage (N₀-N₃). In addition, the T-staging rules include several exceptions and are therefore more extensive than the N-staging rules. This is illustrated by the fact that only the location of the pathological lymph node is different in the N-staging process, whereas for the T-staging, tumor size, presence and involvement differ per stage. Furthermore, to accurately T-stage the tumor, the size is of utmost importance leading to an accuracy score of 0.80 and 0.81 when only finding the accurate tumor size. The additional 0.07 and 0.11 is achieved by setting multiple rules, which is a laborious process.

To increase overall TN-outcome, both T-staging and N-staging processes should be improved. However, it is thought that accuracy for the T-staging is limited, even with finding more synonyms using this single rule-based approach. At this point, changing the rules of the classification process is a tradeoff between improving one rule while

decreasing the outcome of the other. Instead, it may be better to improve T-staging outcomes with machine learning by, for instance, specific training to find difficult-to-extract concepts or match the right context. For example, accuracy may then be improved in case of better identification of gravity-dependent atelectasis, matching uncertainty mentions to the correct concepts or finding specific T4 exceptions.

Furthermore, in a single radiological report often several (pathological) lymph node stations are described. This is beneficial for an NLP algorithm, since, even when a pathological lymph node is missed, another pathological lymph node (in the same level or leading to the same stage) may be picked up by the algorithm, not changing the final outcome. Furthermore, the word “lymphadenopathy” is highly specific for pathological lymph nodes, and so are its modifiers (location, ipsilateral, contralateral or bilateral). Such ‘back ups’ and specific terms are less present in primary tumor staging.

Furthermore, the combination of concepts for “pathological contralateral lymph node” or “enlarged supraclavicular lymph node” is quite specific for the N₃-stage, which in turn allows for better extraction. When a pathological lymph node in this specific location is found, other lymph node stations are of less importance as the highest N-stage is reached. This can be the explanation for the high combined accuracy score of 97.2% in the N₂-stage, in which N₂ harbors the most lymph node levels. The same may be true for the high combined No-stage accuracy score of 98.5%, in which the accurate distinction is to properly match negations and the rules set to not match any of the pathological concepts, enlarged sizes or the word “lymphadenopathy” to the lymph node levels. Although this ‘back up’ may be beneficial for the final results, the algorithm still needs to highlight *all* pathological lymph nodes correctly to increase the accuracy of the radiological report.

However, the abovementioned is not true when only one pathological lymph node is present in a random (non-specific) location. Perhaps this more specific task or option is the reason that relatively many errors are present in the N₁ and N₃-staging group, with overall 0.71 and 0.88 accuracy compared to 0.99 and 0.97 for the No and N₂ stages. In addition, for N₃ nodes the contralateral side needs to be accurately distinguished. Furthermore, several lymph nodes are described in one sentence, which complicates correct matching of contextual information (e.g. uncertainty, negation) even more.

To increase specific N-stage accuracy scores, dependency relations could be used to have a better idea of which contextual property belongs to which lymph nodes, in case multiple lymph nodes are present in a single sentence, this may improve the accuracy score further. In addition, also ML based NLP implementation might be helpful to find specific terms and mentions. This seems less difficult to train than the T-staging because the rules are less difficult. This implementation of ML should be targeted at full matching between the lymph node and the synonyms for the concept pathological, level synonyms and the nodal size as all nodes need these descriptions.

A different approach to increase accuracy, without artificial intelligence tooling and without IT interference, can be through standardization of the report. This standardization step, which is specifically *not* a template or a structured report, represents a set of simple rules on how to report. This can be as simple as stating the size of the tumor or the pathological lymph node directly after the stated concept, perhaps between brackets. A different option is to only give sizes for pathological lymph nodes and tumors, or mention only one lymph node level per sentence. Alternatively, the primary tumor or specific lymph node with all its highlights is reported in one sentence. This way, the set of rules will result in higher accuracy scores. In addition, readability of the reports will improve as well, even without difficult and extensive interventions during the reporting process. The GUI that has been developed that allows real-time analysis and feedback to the reporter may also be beneficial here. As mentioned earlier, this overall quality report enhancing step can also be achieved by ML only, but requires a vast number of reports to train all variants. Also, more annotated data is then needed and, as this is laborious for training purposes, hence not desirable. When looking at the total errors in the training and validation sets, the reporter error group '*reporter*' is responsible for 21.9% of the total errors. These errors are caused by incomplete or inaccurate information, or speech or typing errors when for instance adjusting the report. It is difficult for a rule-based algorithm to find these errors, because staging information can be implicated in the text rather than explicit mentioning, and typos or speech errors occur in many different ways. Even with manually determining classification from the reports it was sometimes difficult to interpret the correct classification. Knowledge of the reporting and staging process - for instance order specific information, overall contextual information, or by knowing what item to

prioritize in wrongly stated concepts or context - made it possible to determine the correct stage. These errors cannot be solved with a rule-based approach and it is questionable whether ML will do better, as these errors may not occur systematically - even in a vast number of reports - hampering the ability of ML to recognize these.

Again, it is very interesting to see whether reporter induced errors can be diminished when the report is staged live and outcomes are displayed using a GUI, as these errors are relatively easy to prevent. From the grand total of 192 scans, 7 staging errors could have been prevented. It seems that an improvement in reporting skills, combined with the implementation of specific ML, would increase accuracy outcomes scores most.

Limitations

One of the limitations of this study is that there were only 192 cases included. Table 1 (Cohort composition of the training and validation set) shows that only a subset of the 32 TN-stages were present in the training and validation set. The relatively few reports included in these groups induced heterogeneity. Furthermore, the outcomes of the additional N-stage were only based on reports from one institution. Future work requires its external validation. In addition, Positron Emission Tomography (PET)-CT reports were not included in this study, in which possibly important tracer uptake information is missed. This is mainly important to exclude enlarged lymph nodes without uptake and include small lymph nodes with tracer uptake.

To complete fully assisted TNM-staging, M-stage is also needed. This is, however, expected to be much more difficult or may be even not feasible at all. For instance, brain metastasis can only be seen with high accuracy on a brain MRI. In addition, also (whole body) PET-CT is used as a screening tool to search for distant metastasis. However, suggested distant metastasis on PET-CT mostly requires additional, specifically targeted imaging to confirm metastasis. As such, only metastasis located in the chest can be found on a staging chest CT, and only those can be staged. In future research, PET-CT reports need to be validated. Merging information from different radiological staging reports is needed for accurate full TNM-staging. Perhaps, such an algorithm can be useful for and applied to oncology staging forms or multidisciplinary meetings.

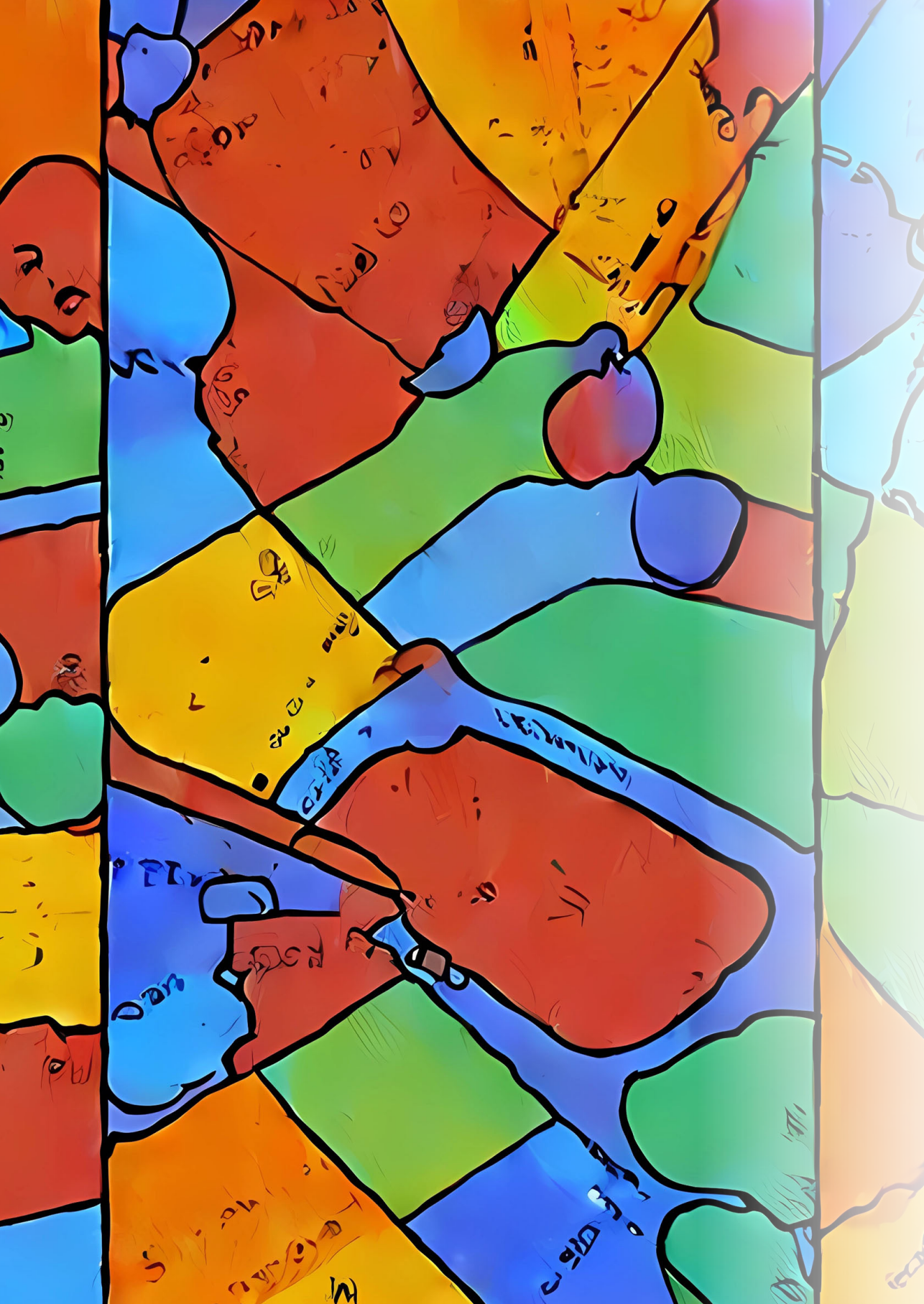
Conclusion

NLP shows its potential in classifying pulmonary oncology from free text radiological reports according to the TNM classification system as both the T and N-stages can be extracted with high accuracy. Integration with machine learning approaches to perform specific tasks should improve accuracy scores even more. However, standardization of the reporting manner and a visual check by the reporter before finalizing the report may be relatively easy implementations in clinical practice to increase accuracy.

References

1. Detterbeck FC, Boffa DJ, Kim AW, Tanoue LT. The Eighth Edition Lung Cancer Stage Classification. *Chest*. 2017;151(1):193-203. doi: 10.1016/j.chest.2016.10.010.
2. Lukaszewicz A, Uricchio J, Gerasymchuk G.. The Art of the Radiology Report: Practical and Stylistic Guidelines for Perfecting the Conveyance of Imaging Findings. *Can Assoc Radiol J*. 2016;67(4):318-321. doi: 10.1016/j.carj.2016.03.001.
3. Goergen SK, Pool FJ, Turner TJ. Evidence-based guideline for the written radiology report: methods, recommendations and implementation challenges. *J Med Imaging Radiat Oncol*. 2013;57(1):1-7. doi: 10.1111/1754-9485.12014.
4. Bosmans JM, Peremans L, Menni M, De Schepper AM, Duyck PO, Parizel PM. Structured reporting: if, why, when, how-and at what expense? Results of a focus group meeting of radiology professionals from eight countries. *Insights Imaging*. 2012;3(3):295-302. doi: 10.1007/s13244-012-0148-1.
5. Ranschaert ER, Morozov S, Algra PR, editors. *Artificial intelligence in medical imaging: opportunities, applications and risks*. Cham (Switzerland): Springer Nature; 2019.
6. Steinkamp JM, Chambers C, Lalevic D, Zafar HM, Cook TS. Toward Complete Structured Information Extraction from Radiology Reports Using Machine Learning. *J Digit Imaging*. 2019;32(4):554-564. doi: 10.1007/s10278-019-00234-y.
7. Pons E, Braun LM, Hunink MG, Kors JA. Natural language processing in radiology: A systematic review. *Radiology*. 2016;279:329- 343. doi: 10.1148/radiol.16142770.
8. Jungmann F, Kuhn S, Tsaur I, Kämpgen B. Natural Language Processing in der Radiologie : Weder trivial noch unerreichbare Magie [Natural language processing in radiology : Neither trivial nor impossible]. *Radiologe*. 2019;59(9):828-832. doi: 10.1007/s00117-019-0555-0.
9. Sorin V, Barash Y, Konen E, Klang E. Deep Learning for Natural Language Processing in Radiology-Fundamentals and a Systematic Review. *J Am Coll Radiol*. 2020 May;17(5):639-648. doi: 10.1016/j.jacr.2019.12.026.

10. Lee SJ, Weinberg BD, Gore A, Banerjee I. A Scalable Natural Language Processing for Inferring BT-RADS Categorization from Unstructured Brain Magnetic Resonance Reports. *J Digit Imaging.* 2020;33(6):1393-1400. doi: 10.1007/s10278-020-00350-0.
11. Zeng Z, Espino S, Roy A, Li X, Khan SA, Clare SE, et al. Using natural language processing and machine learning to identify breast cancer local recurrence. *BMC Bioinformatics.* 2018;19(Suppl 17):498. doi: 10.1186/s12859-018-2466-x.
12. Lou R, Lalevic D, Chambers C, Zafar HM, Cook TS. Automated detection of radiology reports that require follow-up imaging using natural language processing feature engineering and machine learning classification. *J Digit Imaging.* 2020;33(1):131-136. doi: 10.1007/s10278-019-00271-7.
13. Abdulsalam AKAAI, Garvin JH, Redd A, Carter ME, Sweeny C, Meystre SM. Automated extraction and classification of cancer stage mentions from unstructured text fields in a central cancer registry. *AMIA Jt Summits Transl Sci Proc.* 2018;2017:16-25.
14. Spandorfer A, Branch C, Sharma P, Sahbeea P, Schoepf UJ, Ravenel JG, et al. Deep learning to convert unstructured CT pulmonary angiography reports into structured reports. *Eur Radiol Exp.* 2019;3(1):37. doi: 10.1186/s41747-019-0118-1.
15. Pruitt P, Naidech A, Van Ornam J, Borczuk P, Thompson W. A natural language processing algorithm to extract characteristics of subdural hematoma from head CT reports. *Emerg Radiol.* 2019;26(3):301-306. doi: 10.1007/s10140-019-01673-4.
16. Yetisgen-Yildiz M, Gunn ML, Xia F, Payne TH. A text processing pipeline to extract recommendations from radiology reports. *J Biomed Inform.* 2013;46(2):354-62. doi: 10.1016/j.jbi.2012.12.005.
17. Nobel JM, Puts S, Bakers FCH, Robben SGF, Dekker ALAJ. Natural Language Processing in Dutch Free Text Radiology Reports: Challenges in a Small Language Area Staging Pulmonary Oncology. *J Digit Imaging.* 2020 Aug;33(4):1002-1008. doi: 10.1007/s10278-020-00327-z.



Chapter 7:

Natural Language Processing Algorithm used for Staging Pulmonary Oncology from Free-text Radiological Reports: including PET-CT and validation towards clinical use

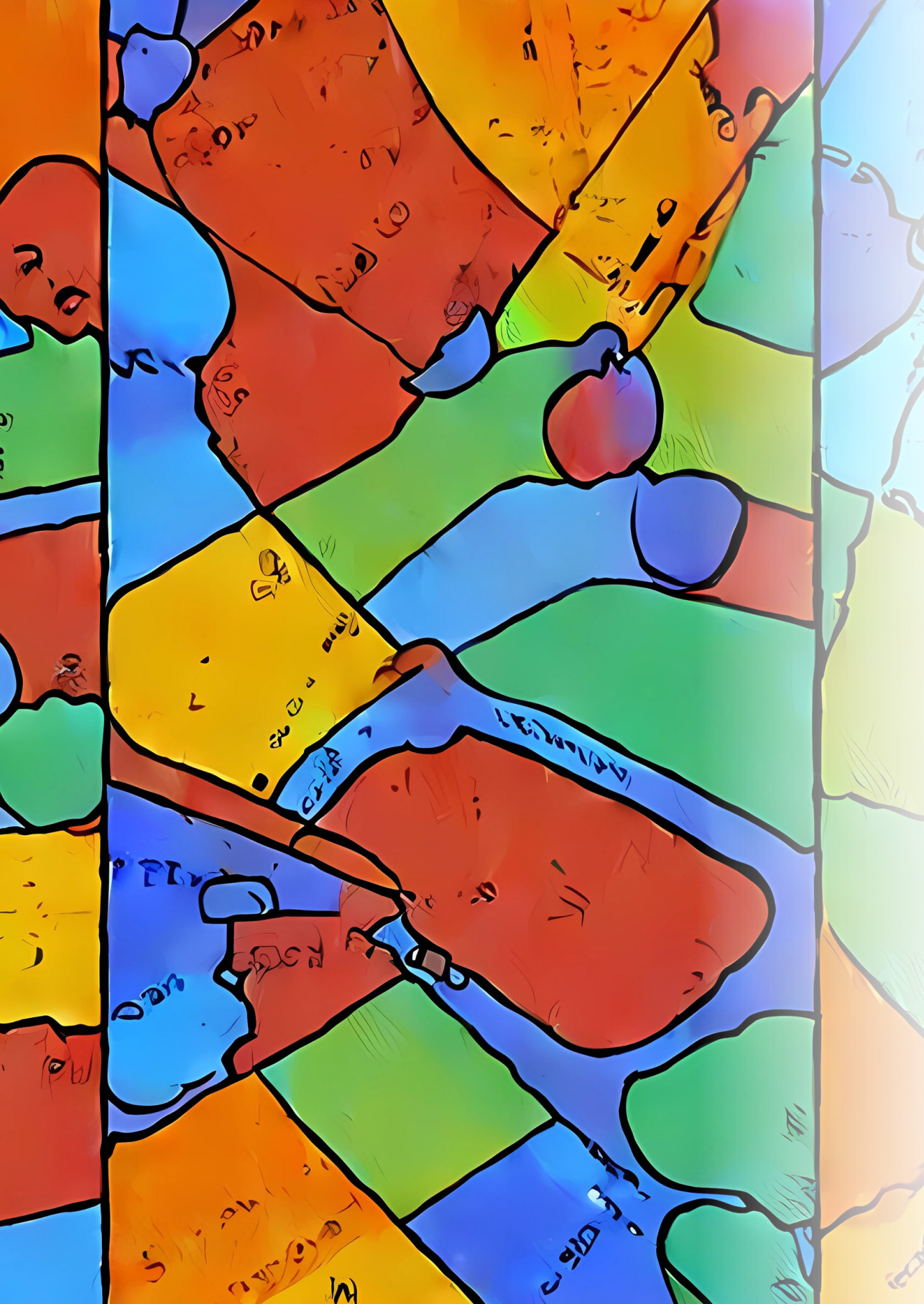
EMBARGOED

J. Martijn Nobel, Sander Puts, Jasenko Krdzalic, Catharina M.L. Zegers,

Marc B.I. Lobbes, Simon G. F. Robben, André L. A. J. Dekker

Submitted to Journal of Digital Imaging (2023)

GENERAL DISCUSSION
AND SUMMARY



Chapter 8:

General discussion



The overall aim of this thesis is to better understand how to improve reporting in radiology. Because the content of the radiological report and its format vary widely, this is considered as an important issue to improve. Part One of this thesis focuses on structured reporting (SR) in radiology, as it is believed to be a promising tool for improving the reporting process. In Part Two, the efforts made on the implementation of Natural Language Processing (NLP) in the radiological reporting process as an additional tool have been explored. Both SR and NLP can be used in the radiological reporting process as possible solutions to improve the radiological report and this thesis aims to discover their potential.

In order to address this overall goal, several research aims were set:

PART I: Radiological report and reporting process

- Explore what structured reporting entails and what its definition is
- Review the literature on structured reporting to explore its current implementation and to determine the level of evidence.

PART II: Text mining and Natural Language Processing

- Explore how free text mining and Natural Language Processing (NLP) can extract the T-stage (TNM classification in primary lung carcinoma) from a free text report, in a Dutch setting.
- Assess how this rule-based NLP T-staging algorithm can be translated and trained in an English setting
- Extend the existing Dutch NLP T-staging algorithm towards a TN-staging algorithm
- External validation of the Dutch NLP TN-staging algorithm
- Extend the existing Dutch NLP TN-staging algorithm with PET-CT functionality

In the following paragraphs the main findings and future perspectives will be explained and discussed.

PART I: Radiological report and reporting process

Since the beginning of reporting in radiology only little has changed in the way of reporting [1-4]. The reporting process starts with a clinical question that the radiologist answers in the radiological report after the examination has been performed by the technician. Several attempts and guidelines to improve the radiological report can be mentioned, especially in the last decades [2, 5-17]. One of the last attempts for improvement is the ongoing call for standardization and the even more recent move towards structured reporting. Especially structured reporting is widely promoted by international radiological societies in an attempt to increase the accuracy and readability of the radiological report [18-19] (Fig. 1).

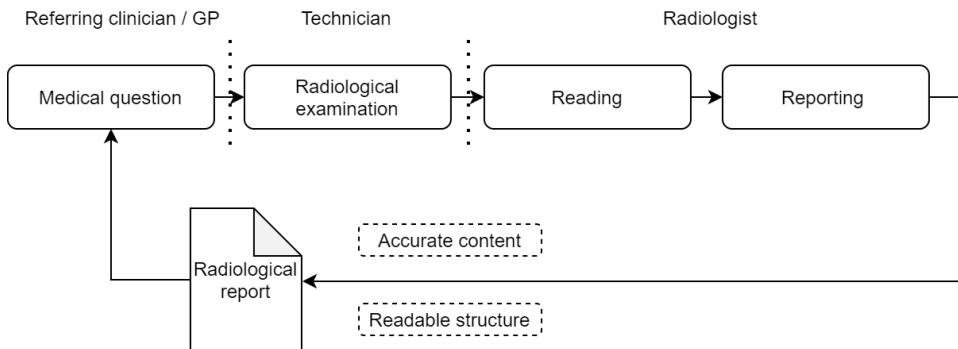


Figure 1. The basic radiological reporting process.

In **Chapter 2** the process of understanding structured reporting in radiology is described as its meaning and interpretation are not clear in scientific literature. The main issue is the lack of a proper definition and the overlap with the interpretation of the meaning of standardized reporting. Therefore, this chapter provides an overview of the different interpretations of structured reporting in literature as numerous characterizations of the term structured reporting are known. Only when its definition is clear, can it be evaluated scientifically, in order to facilitate evidence-based implementation. This chapter suggests a more precise definition of SR that facilitates a better categorization of the heterogeneous literature.

After reading the extensive, but chaotic literature, it was clear that the definition of structured reporting has become ambiguous and often confused with standardization. As described in this chapter we propose that “standardized reporting is supposed to be a means of streamlining the medical content of a radiological report”, whereas “structured reporting is supposed to be the use of an IT-based means of importing and arranging medical content in the radiological report”. A further subdivision of SR was necessary to create more structure in this topic in which SR level 1 (structured layout) was distinguished from SR level 2 (structured content). SR level 1 presents the report findings in a strict, predefined order, creating and maintaining uniformity (e.g. templates or “head to toe” reporting), whereas SR level 2 is the manner in which the medical content is arranged and displayed in the report (e.g. drop down menus or point-and-click systems). With these proposed definitions a clear distinction between standardized reporting and different levels of SR is presented and proper research can be facilitated.

We realized that standardized reporting and SR might both be important in the radiological reporting process and that, based on the difficulties in extracting clear definitions, they seem to be highly intertwined. However, each has its own strength and field of application. In order to evaluate the value of these individual components it is not only necessary to know their definition, but it is also important to know how these tools can support the reporting process.

By using the aforementioned (re)defined definitions we were able to classify articles found in literature on structured reporting in order to determine the level of evidence of these articles. This process has been described in **Chapter 3**. To search for the level of evidence of SR in radiology reporting, a narrative systematic review was performed. In addition, an overview was created of the current status of SR in radiology. In total, 8561 articles were found, resulting in 63 relevant papers (44 SR level 1 and 19 level 2). Only one study performed a double cohort study with randomized trial design and scored the highest level of evidence. The overview of the current status of SR showed a bonanza of different study protocols, research questions and outcomes, and underlines the difficulties of understanding the subject and its meaning. As a consequence, it is not

clear whether SR is truly beneficial in the radiological reporting process or that it just is a means to structure its content.

When focusing on the study protocols, especially the content of the radiological report seems an important goal in enhancing the radiological report by using SR. Multiple studies try to enhance the content of the report with standardized phrases, items or some sort of classification system. This is an interesting finding, especially when it is supposed that standardized reporting should increase report standardization, and not SR. In these cases, SR is often used to implement standardized reporting. Perhaps this may explain the difficulties in finding clear definitions for both concepts in literature as both are highly intertwined – at least in clinical research.

Another observation is that the solution to improve the radiological report is often found in a mandatory manner, stating specific items or by enforcing the reporter to use a specific format or classification system. This again is an interesting finding as it seems that we need SR to structure and ensure report completeness. Why not use a piece of paper or a simple paper or online guideline in order to improve report content? No, instead we need (expensive) SR to be more precise. Of course this is somewhat overstated, but the main question is why standardized reporting alone is not sufficient. This is probably because of the lack of discipline to follow these guidelines results in persisting heterogeneity in the way of reporting among different reporters. SR is helpful and at the same time this facilitates standardization.

This statement can be highlighted by focusing on the only high-level evidence study included in this review that had the only negative outcome when implementing SR. This study compared free text with SR, without a standardization step during the study, and thereby only changing the way of reporting. Without standardization (because this was done earlier in the clinical setting), the outcome was not beneficial for SR and the outcome measurement even worsened. On the contrary, the lower level studies, that incorporated some sort of standardization during the implementation of SR, showed beneficial or equal outcomes when comparing free text with SR. This again shows that standardization is a very important factor, and perhaps the main factor, to improve the radiological report. In Fig. 2 the reporting process is depicted once more, and standardization and SR are added. As standardization seems to be at least as important or even more important than SR, standardization is placed before SR.

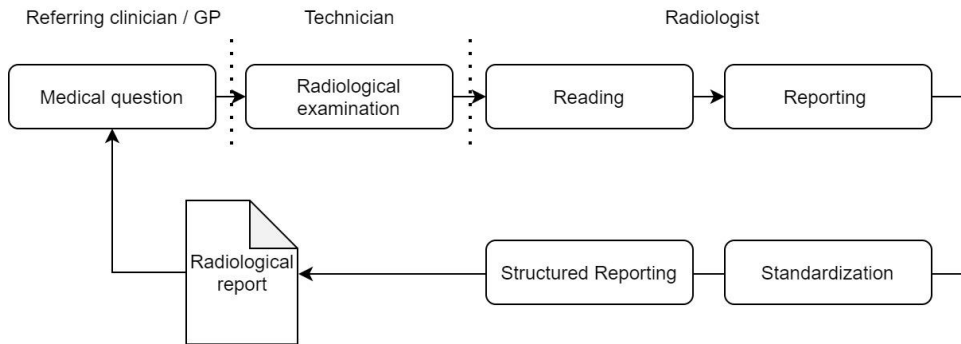


Figure 2. The radiological reporting process, structured reporting and standardization added.

In addition, this overview shows a wide variety in SR techniques used and outcomes, but almost all studies changed several things (e.g. standardization) at the same time as they implemented some sort of SR, which hampers the reliability of the outcomes. In addition, it is possible that some sort of SR only favors easy to make yes/no decisions, rather than more difficult descriptive ones. Thereby, the effect of a particular question or report description can be in favor of one of the two different SR subtypes, whereas it does not match other reporting indications. For instance, the description of a thyroid nodule on ultrasound – which is a fairly easy task with few options to describe – can have a different performance using the same SR method than the description of the location of a brain metastasis on MRI – which has a large description diversity and many anatomical locations. This shows that the value of SR might be organ system-, modality- or even question specific and that there is no “one size fits all”. In addition, none of these studies performed a comparative trial to see which type of SR was most efficient in a particular case. Overall, we can conclude that the level of evidence for SR is rather low despite its promotion by international societies.

Future perspectives

To further elaborate whether SR is indeed a facilitator for standardization, proper and focused research is necessary. Controlled trials with only changing one variable are key. In addition, comparison studies should be made modality-specific, body part specific and perhaps also question specific, as SR is thought to be highly case-specific. The

reason for this high case-specificity is the wide variety in task-complexity as already explained. In the same way we can imagine that also yes/no questions are better depicted with SR than for instance location specific items, at least when it is necessary to point out the exact location of the affected organ. Of course this is not strict, because by using yes/no questions the reporter is forced to choose from the given options only. This might be the reason that reporting and data systems (RADS) are used in many included studies trying to validate SR, as it is standardized already, and can be answered with yes/no already. Nevertheless, when SR is being added to the reporting process using standardized items or standardized language, this new reporting manner should be compared using the same conditions as before the implementation of SR. Another perspective is that SR reports provide structured data, and because structured data is findable data it is easy to be reused for all kind of research, education and quality purposes.

In part II we further elaborate on NLP as a possible substitute for SR. Text mining and NLP as artificial intelligence (AI) tools might be used to accurately process large quantities of text data. When the radiological report consists of SR or SR elements this would highly increase the field of data-mining in radiology, because data is becoming findable. This advantage should also be considered when implementing SR in clinical practice, because a free text report is less easy to process.

A different opportunity when using NLP is that it can help real-time with standardizing and structuring the report when it is implemented in the reporting process.

Overall, when focusing on radiology, a combination of NLP and radiomics (text and image analysis) might be very promising. This is mainly because the radiological report content is still the golden standard.

PART II: Text mining and Natural Language Processing

As SR can facilitate text mining by using NLP, NLP can also function as counterpart for SR as it can mine (unstructured) free text. Of course it is better and easier to use structured data for data mining, but NLP can be used to mine free text radiology reports as well. This is of particular interest because SR is not yet widely implemented in clinical practice and perhaps will not be the solution we are aiming for. Therefore, it is important to search for different solutions that can improve the radiological report and/or its reuse.

As a structured report made with SR should facilitate better report content and might eventually lead to a better radiological report, NLP facilitates text mining and can thereby be a possible solution for enhancement of the radiological report as well. NLP is used already in radiology by processing free text radiological reports in several different settings [20-22]. Implementation of NLP aims at the extraction of specific descriptive entries in the free text radiological report that can be used for further analysis. Examples are fracture or pneumonia detection where outcomes are used for advising additional examination or recommendations for antibiotics [23-25]. NLP is also used in oncology, when extracting information out of pathology reports, medical records or free text radiology reports for case identification, staging or outcome quantification [20]. In this way, specific report information is used in which artificial intelligence embodied in NLP can use the information embedded in the radiological report.

In Part II we investigated whether AI by means of NLP can be used to extract the Tumor and Nodal status according to the Tumor Node Metastasis (TNM) classification [26] out of free text radiological chest (PET) CT reports in order to assess the potential of NLP in radiology reporting. When it is possible to extract specific data necessary for lung tumor staging out of the free text report, we can use it as structured data for data mining purposes. In addition, when we are able to perform these data mining tasks during the reporting process or just before finishing the report, it is possible to enhance the radiological report as well. In this way NLP can function as substitute for SR (Fig. 3)

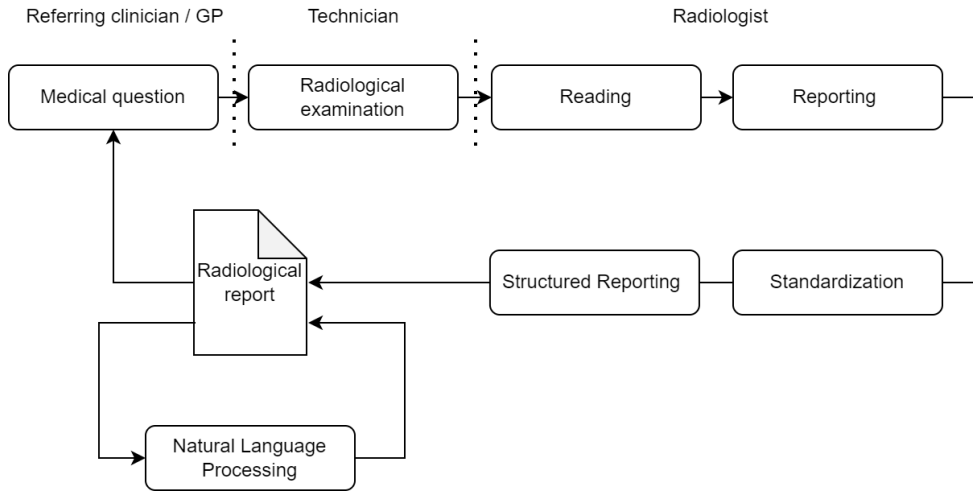


Figure 3. The radiological reporting process including structured reporting, standardization and Natural Language Processing.

Lung carcinoma staging was used for this research question, because it is the most common cause of oncological death worldwide and because radiological imaging has an important contribution in the staging process [27].

In **Chapter 4** the process of applying NLP to this task was described to explore the capability of extracting T-stage of lung carcinoma out of Dutch free text CT thorax radiological reports. A rule-based CT-T algorithm was built, trained and validated on respectively 47 and 100 cases to extract concepts stating something about the items *tumor size*, *presence* and *involvement*, necessary for lung carcinoma staging according to the TNM classification system. Due to the extensiveness of the rules of the TNM classification system, it was not possible to use machine learning, because we lacked a large quantity of specific staging data necessary to train this algorithm. On the contrary, we found that making a rule-based algorithm was quite easy, because of the strict rules of the TNM classification. As a consequence, it was not necessary to use ML for finding these rules. By doing so we learned that the process of building this algorithm is very specific and that field expertise is of great benefit. However, finding the correct concepts and synonyms and to correlate them to the correct context is more difficult. Especially

because of the free format and because there were no widely used reporting rules, it was difficult to decide and find appropriate rules that were able to accurately stage the reports. Even more, recognition of speech errors or omissions as well as the right correlation of numbers or negations to a specific concept were difficult tasks. In the end overall accuracy of the Dutch pulmonary carcinoma T-staging algorithm was 0.83 in the training set and 0.87 in the validation set, showing its potential feasibility and that the rule-based approach was quite successful.

Chapter 5 describes the process of translating the rule based Dutch lung carcinoma CT-T algorithm into English. Because the TNM and the content of the radiological report should be – apart from the text of course – roughly the same, this was a good opportunity to validate this algorithm in English. The algorithm was trained and validated on English free text radiology (PET-)CT reports used for primary lung carcinoma staging. 200 reports were used for training and 225 for validation, resulting in an accuracy score of 0.89 in the training and 0.84 validation set. Systematized Nomenclature of human MEDicine-Computed Tomography (SNOMED-CT) terms were used for the translation of the used concepts from Dutch to English. Vocabularies like SNOMED-CT are supposed to be useful in structuring text by coding concepts and its synonyms as well as their connecting word families and are available in different languages. However, in our case using only the same words according to the SNOMED-CT codes was insufficient to translate all synonyms and, because several Dutch concepts did not match the meaning of the English counterparts, it was still necessary to search for English synonyms in the radiological report. In this training period a graphical user interface (GUI), called MEDSTRUCT-NLP (Fig. 4) was built for to visualize the results of the algorithm by highlighting the extracted concepts and its modifying context. By doing so we were able to see what the algorithm was doing and what was (falsely) matched. During this testing process the GUI was found to be so efficient as a structuring tool that we decided that such a tool could be useful in the radiological reporting process as well. Because the GUI highlights the stated free text in a structured format, this can function as a visual check whether all necessary items are stated in the report. Especially when this check can be done real-time or just before finishing the radiological report, this will enhance the radiological report. In addition, when

information is added by processing the structured entries it can add additional value to the radiological report, as in this case with the addition of the TN-classification.

Figure 4. Graphical user interface MEDSTRUCT-NLP [28].

To add more functionality to the Dutch free text lung carcinoma T-staging algorithm, an N-stage classifier was added to the prior T-stage classifier, resulting in a TN-CT algorithm. The process of implementation, training and validation was described in Chapter 6. A set of respectively 95 and 97 CT scans was used for training and validation and the N-stage accuracy scores were 0.96 and 0.92. The TN-stage accuracy scores were 0.84 and 0.85 for this training and validation set. As this was foreseen as a difficult task the outcomes were outperforming our expectancies. The algorithm was programmed to match the location/laterality of the tumor – which was already known from the T-part of the algorithm – with the pathologic lymph node levels. Then the location of the lymph node with the highest N-score is depicted as final N-score. The high performance scores might be explained by the fact that the description of the affected lymph node

levels were most often repeated in the findings section and only little variation in description of the lymph nodes is found in the report. The same was true for the laterality annotation of the primary tumor.

In the clinical staging process of lung carcinoma also a Positron Emission Tomography – Computed Tomography (PET-CT) scan is used which adds metabolic information about the uptake of fluorodeoxyglucose (FDG) by a tumor or lymph node in addition to the anatomical information provided by the CT scan. **Chapter 7** describes the final study on how Dutch PET-CT reports in lung carcinoma tumor staging can be staged with this NLP-based tool. To enable this, the Dutch CT-TN algorithm was extended with a metabolic layer to result in a PET-CT-TN algorithm. In total 63 (24 CT, 39 PET-CT) and 100 (41 CT, 59 PET-CT) radiological reports were included for training and validation. TN accuracy scores were respectively 0.73 and 0.62 for the training and validation set. In addition, and because this study was performed in a different Dutch hospital, a subgroup analysis could be performed to externally validate the CT-TN algorithm. This TN accuracy score was 0.72. Both outcomes show a lower performance score and are somewhat disappointing. Especially the lower external validation score was not expected. However, when looking more into detail to the outcomes it seems that location specific vocabulary or stating of certainty levels differ more between different hospitals than expected. It was also remarkable that the PET-CT-TN performance was worse in comparison with the CT-TN performance, because we expected that adding metabolic information would increase accuracy instead of decreasing it.

When combining the outcomes of chapter 4-7 it is seen that a rule-based approach of extracting the Tumor and Nodal status from free text radiological reports show mixed results. The goal of this explorative mission was to look for the boundaries of the application of NLP. This rule-based approach was chosen because of the high difficulty of the task and because not many reports were present to allow for the use of machine learning (ML). ML usually needs a lot of data and, depending on the task difficulty, even more data to extract the staging rules as well as to be capable of finding the right items necessary for staging. The rule-based approach was therefore chosen.

We showed that this rule-based NLP approach is good enough to start with and understanding the way the algorithm should function is pivotal to train the separate steps of the application. A notable finding was that this rule-based algorithm was trained fairly easily, considering the effort that would have to be made when only using ML. In addition, by using a rule-based approach the algorithm can be easily adjusted when necessary. As a result, this rule-based algorithm is dependent on the combination of task specific knowledge and NLP, rather than AI alone. Simple adjustments by using ML do not exist, because the algorithm should be trained all over again.

It was interesting to see that outcomes were roughly the same in the English setting, but remarkable that using SNOMED-CT alone for translation of used concepts was not sufficient. This shows that the SNOMED-CT translation terms do not fully match vocabulary meaning and use between languages. However, a small training set allowed us to train the algorithm to be capable of reaching shown accuracy scores. This shows a way to use this approach in different languages (and perhaps for different tasks) as well. However, the PET-CT-TN algorithm accuracy scores and the external validation of the CT-TN algorithm are underperforming compared to the CT-TN algorithm and the primary training center. This is probably caused by differences in reporting between different institutions in content and vocabulary as well as differences in focus considering the metabolic versus anatomical focus, when comparing PET-CT and CT examinations. In addition, several different errors do occur due to differences between used vocabulary, the description of its certainty, their dependencies as well as finding the correct report section to extract correct information.

Future perspectives

The most important issue is to construct a PET-CT-TN algorithm that is institution independent or at least can adapt itself to different institutions. More reports from different institutions are necessary to achieve this. In addition, the use of ML may increase the accuracy of finding the right items, their certainty and dependencies. Especially the focused use of ML is considered to be an important addition to the existing algorithm, as the rule-based approach is already working decently. An ML application focused on recognizing synonyms and comparatives as well as finding the right sections might enhance, in specific cases, the accuracy of NLP. However, only

larger topic-specific datasets allow for more ML/AI driven solutions and specific oncological datasets are scarce. This hampers a wider implementation of ML in this field. In addition, also non topic-specific tasks can be trained on regular datasets, for instance looking for non-tumor dependent atelectasis in non-oncological staging chest CT reports – as tumor dependent atelectasis is a difficult-to-extract entity. Then it will be possible to strengthen difficult parts of a rule-based algorithm with ML. Another example is to correctly match sizes (or diameters) with their corresponding concept. This does not need to be a tumor, but can also be a node, a cyst, a pleural effusion or something else, as long as the ML focuses on the interaction between the size and the concept with its context. Again, particular tasks can be trained without specific datasets and this approach seems to be promising for training tasks on smaller datasets. We think such a modular approach is necessary for all future NLP algorithms, whereby (pre)processing tasks can be more sensitive, leading to better annotation and understanding of the concepts and their context.

Furthermore, a GUI, such as we developed, seems to be a great opportunity to increase the reporting content as it can, like SR, provide a tool in which reporting content can be checked real time while a radiologist is dictating which will result in better reporting content. Of course, this again is an IT/AI based tool and perhaps simply using a guideline will enhance the radiological report as well, but this tool leads to less or even no interfering into or adding to the reporting process which is not the case for SR. The report can still be a free text report and structuring and classification processes can be done during reporting without mandatory clicking or interrupting the used way of reporting. This shows that NLP can be a counterpart for SR and future research should focus on the interaction between these two entities.

Concluding remarks

The efforts made on the improvement of the radiological report over last decades did not yet result in a sustainable solution for the problems encountered. Three different solutions are being discussed in this thesis that can increase accurate report content and that may lead to a more accurate content and readable structure:

1. Standardized reporting
2. Structured reporting
3. Natural Language Processing

First, standardization seems the most likely first step to take, as it seems to be the easiest solution to increase report content without great changes to the reporting process. Many standardization tools have already been developed and can be implemented with or without IT-implementation. When the content of the radiological report is standardized, the second step is looking for evidence for different types of SR. This is probably a difficult task, as it seems that its use might be highly case-specific. Therefore, the hypothesis is that implementation is case-specific and that the type of SR tool and its level should be adjusted accordingly. Finally, NLP can also assist the reporter to be more complete by adding or unlocking information about for instance the TNM classification, as shown in this thesis.

These three steps are separate entities with different aims and outcomes. However, it is shown that standardization and SR are highly intertwined and that standardization might be an influencing factor in the use of NLP tooling as well. Therefore, enhancing the radiological report and reporting process will be multifactorial as shown in Fig. 5.

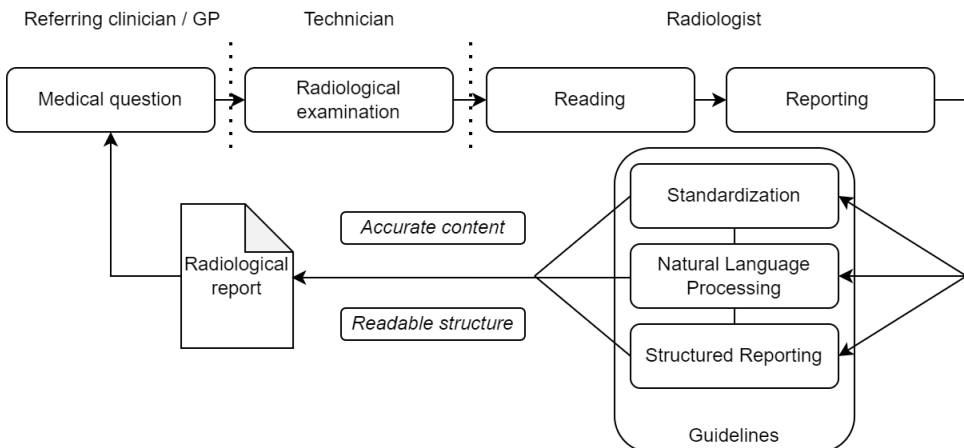


Figure 5. Schematic visualization of efforts made in radiological reporting process to improve the radiological report and as discussed in this thesis.

It is also interesting when zooming out from NLP use in radiology reporting towards NLP usage in the Electronic Health Record (EHR). Especially, in times of increasing numbers of burnout and stress among physician and other healthcare employees it is necessary to look for the impact of health information technology [29-31]. Medical healthcare is facing an ever-growing pile of paperwork as well as an increase in administrative burdens. Structured reporting and above all NLP can potentially be helpful in attacking this datafication monster.

In the EHR, large amounts of unstructured data are stored and, when available for (re)use, it can have a great impact on modern healthcare. Transferring information from the EHR to the radiological report and vice versa can increase workflow for the referring clinician and reporter and/or can enhance logistics when a follow-up or additional study needs to be performed. A more specific risk stratification can be made when clinical information can be used in for instance cancer staging, or data linking can result in better survival prediction. In this way, linking data can enhance the progress made in digital healthcare. Nowadays, extraction and registration of data is mainly done manually, but when technical solutions, like for instance NLP, take over some tasks as by using a GUI or staging algorithm, this can be a significant relief.

A different approach to enhance radiology and medicine, is when we add image mining (e.g. radiomics) to the solutions made on the improvement of the radiology report as shown in this thesis. After all, AI is already being used on a large scale in image mining in radiology, but information extracted from the image should be matched and compared with the textual information and implemented in the final radiology report. As such, the radiological report still is seen as the golden standard. Examples of the use of AI and/or radiomics can be a particular size or fracture type that is inserted into the report, or perhaps a suggestion for a particular tumor class based on the images. To go further, it would be a real progress and goal of future research to add information about follow-up into the report or directly plan an additional examination or follow-up appointment. To go even further, it should be possible to combine the information of the radiological report, the radiological examination to the information stored in the EHR to allow for even wider use of available data in order to facilitate the progress in the healthcare process. For instance, integration and combination of radiology data

elements with other key clinical parameters (e.g. laboratory results), leading to an integrated and precise diagnosis, and beyond, to computer-assisted clinical decisions. Finally, in the current quest for improving the radiological report, an approach has to be found that assists the reporter, but preferably without interfering too much with the reporting process and personal reporter preferences. This will always be a tradeoff, compromising one thing or the other. One of the most interesting things for future improvement is the use of an (interactive) GUI to assist reporting. As we presented in this thesis, a GUI can enhance the reporting process by showing the report content and can assist in it being more complete. As such, a GUI can assist in streamlining the medical content and its structure, whether the reporting process is enhanced with standardization, SR, NLP or a combination (Fig. 6). The solution that is most interesting is a system that checks (real-time) and structures the report content, without interfering in the reporting process. This can be as simple as (re)placing information in the right section or as complicated as adding information from a different external system like for instance the EHR or from the image directly. A pop-up system or content-dependent suggestions made by algorithms used by the GUI can therefore enhance the radiological report on several levels.

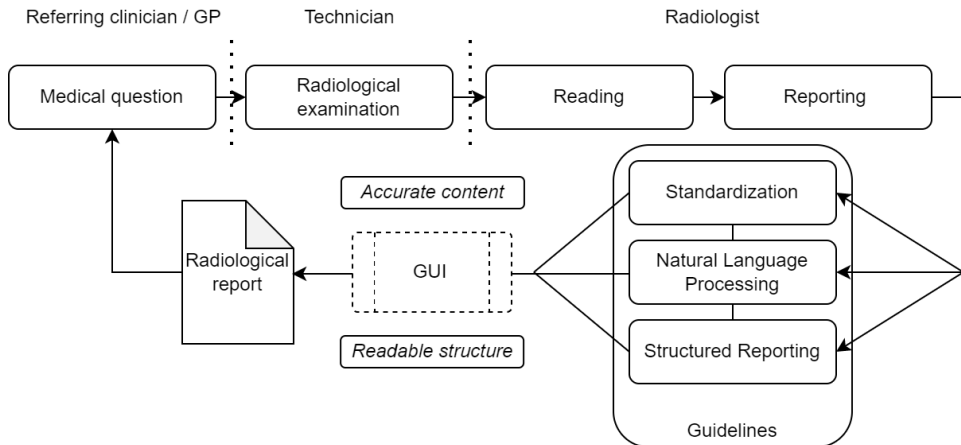


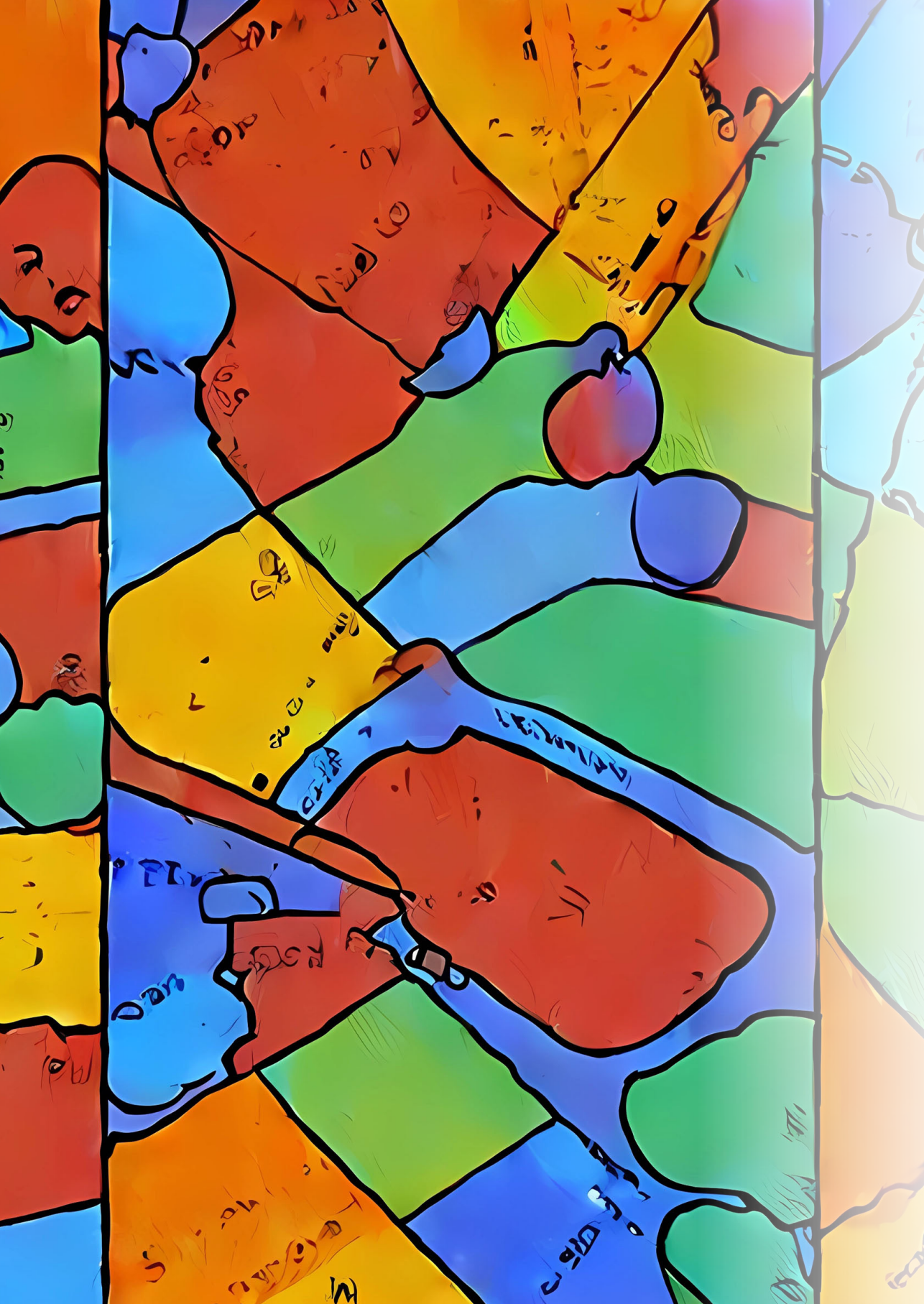
Figure 6. The GUI as facilitator for the implementation of efforts made in the radiological reporting process.


References

1. Wallis A, McCoubrie P. The radiology report - are we getting the message across? *Clin Radiol*. 2011;66(11):1015-22. doi: 10.1016/j.crad.2011.05.013.
2. Langlotz CP. The radiology report: a guide to thoughtful communication for radiologists and other medical professionals. CreateSpace Independent Publishing Platform; 2015.
3. Brady AP. Radiology reporting - from Hemingway to HAL? *Insights Imaging*. 2018;9:237-246. doi: 10.1007/s13244-018-0596-3.
4. Reiner BI, Knight N, Siegel EL. Radiology reporting, past, present, and future: the radiologist's perspective. *J Am Coll Radiol*. 2007;4(5):313-9. doi: 10.1016/j.jacr.2007.01.015.
5. European Society of Radiology (ESR). Good practice for radiological reporting. Guidelines from the European Society of Radiology (ESR). *Insights Imaging*. 2011;2(2):93-96. doi:10.1007/s13244-011-0066-7.
6. Grieve FM, Plumb AA, Khan SH. Radiology reporting: a general practitioner's perspective. *Br J Radiol*. 2010;83(985):17-22. doi: 10.1259/bjr/16360063.
7. Recommandations générales pour l'élaboration d'un compte-rendu radiologique (CRR). *J Radiol*. 2007;88(2):304-6. doi: 10.1016/S0221-0363(07)89822-2.
8. American College of Radiology. ACR practice guideline for communication of diagnostic imaging findings [Internet]. Reston: American College of Radiology; 2005 [cited September 2020]. Available from <https://www.acr.org/-/media/acr/files/practice-parameters/communicationdiag.pdf>
9. The Royal College of Radiologists. Standards for the Reporting and Interpretation of Imaging Investigations [Internet]. London: The Royal College of Radiologists; 2006 [cited September 2020]. Available from https://www.rcr.ac.uk/sites/default/files/bfcro61_standardsforreporting.pdf
10. Siström CL, Langlotz CP. A framework for improving radiology reporting. *J Am Coll Radiol*. 2005;2:159e67. doi: 10.1016/j.jacr.2004.06.015.
11. Reiner BI, Knight N, Siegel EL. Radiology reporting, past, present, and future: the radiologist's perspective. *J Am Coll Radiol*. 2007;4(5):313-9. doi: 10.1016/j.jacr.2007.01.015.
12. European Society of Radiology (ESR). ESR concept paper on value-based radiology. *Insights Imaging*. 2017;8(5):447-454. doi: 10.1007/s13244-017-0566-1.
13. Kahn CE Jr, Langlotz CP, Burnside ES, Carrino JA, Channin DS, Hovsepian DM, et al. Toward best practices in radiology reporting. *Radiology*. 2009;252(3):852-6. doi: 10.1148/radiol.2523081992.
14. Lukaszewicz A, Uricchio J, Gerasymchuk G. The Art of the Radiology Report: Practical and Stylistic Guidelines for Perfecting the Conveyance of Imaging Findings. *Can Assoc Radiol J*. 2016;67(4):318-321. doi: 10.1016/j.carj.2016.03.001.

15. Reiner BI. The challenges, opportunities, and imperative of structured reporting in medical imaging. *J Digit Imaging*. 2009;22(6):562-8. doi: 10.1007/s10278-009-9239-z.
16. Hall FM. Language of the radiology report: primer for residents and wayward radiologists. *AJR Am J Roentgenol*. 2000;175(5):1239-42. doi: 10.2214/ajr.175.5.1751239.
17. Jacoby J, Ayer R, editors. Frameworks for radiology reporting. London: Taylor and Francis; 2009.
18. European Society of Radiology (ESR). ESR paper on structured reporting in radiology. *Insights Imaging*. 2018;9(1):1-7. doi: 10.1007/s13244-017-0588-8.
19. Radiological Society of North America. RadReport template library [Internet]. Oak Brook (IL): Radiological Society of North America; 2020 [cited 15 Dec 2020] Available from <https://radreport.org>
20. Yim WW, Yetisgen M, Harris WP, Kwan SW. Natural Language Processing in Oncology: A Review. *JAMA Oncol*. 2016;2(6):797-804. doi: 10.1001/jamaoncol.2016.0213.
21. Pons E, Braun LMM, Hunink MGM, Kors JA. Natural language processing in radiology: A systematic review. *Radiology*. 2016;279:329- 343. doi: 10.1148/radiol.16142770.
22. Sorin V, Barash Y, Konen E, Klang E. Deep Learning for Natural Language Processing in Radiology-Fundamentals and a Systematic Review. *J Am Coll Radiol*. 2020;17(5):639-648. doi: 10.1016/j.jacr.2019.12.026.
23. Do BH, Wu AS, Maley J, Biswal S. Automatic retrieval of bone fracture knowledge using natural language processing. *J Digit Imaging*. 2013;26(4):709-713. doi: 10.1007/s10278-012-9531-1.
24. Fiszman M, Chapman WW, Aronsky D, Evans RS, Haug PJ. Automatic detection of acute bacterial pneumonia from chest x-ray reports. *J Am Med Inform Assoc*. 2000;7(6):593-604. doi: 10.1136/jamia.2000.0070593.
25. Fiszman M, Chapman WW, Evans SR, Haug PJ. Automatic identification of pneumonia related concepts on chest x-ray reports. *Proc AMIA Symp*. 1999:67-71.
26. Brierley J, Gospodarowicz MK, Wittekind C, editors. TNM classification of malignant tumours. 8th ed. Chichester: John Wiley & Sons Inc; 2017.
27. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J Clin*. 2021;71(3):209-249. doi: 10.3322/caac.21660.
28. Gigase M, Schoonbrood I. Integrating the Healthcare Enterprise (IHE). Handreiking Het transmurale MDO Mammacarcinoom vormgegeven op basis van nationale en internationale standaarden [Internet]. Veenendaal (The Netherlands); 2021 [cited 27 August 2022]. Available from https://ihe-nl.org/wp-content/uploads/2021/05/IHE_MDO_en_Addendum_17_mei_2020_StatusDefinitief.pdf
29. Gardner RL, Cooper E, Haskell J, Harris DA, Poplau S, Kroth PJ, et al. Physician stress and burnout: the impact of health information technology. *J Am Med Inform Assoc*. 2019;26(2):106-114. doi: 10.1093/jamia/ocy145.

30. Top challenges facing physicians in 2021: Administrative burdens and paperwork. *Medical Economics*. 2021;98(1):6-8.
31. Visser MR, Smets EM, Oort FJ, De Haes HC. Stress, satisfaction and burnout among Dutch medical specialists. *CMAJ*. 2003;168(3):271-5.





Chapter 9: Summary

This thesis titled “The Radiological Report: *A compromise between Structured Reporting and Natural Language Processing*”, describes in two parts the search for how to improve the radiological report and the radiological reporting process.

PART ONE: Structured reporting in radiology

In the first part, the application of structured reporting (SR) is discussed. Although efforts are encouraged by international radiological organizations, it has not been sufficiently investigated whether SR actually improves the radiological report. This is mainly because in these studies standardization is combined with SR. In many studies, standardization is enforced by structured reporting; for example by requiring the reporter to fill in fields or to choose from a specific (drop-down) menu. However, it seems that particularly this standardization improves (the content of) the reports and that the influence of SR itself is still questionable. In particular, the lack of research into the different forms of SR means that there is still a lot of work to be done before drawing conclusions. What does seem to work is the mandatory nature of implementation of standardization. However, should this be done with SR?

PART TWO: Natural Language Processing

This second part examines how NLP can be used as a counterpart to SR in the search for improvement of the radiological report. NLP is a form of Artificial Intelligence (AI) that can search texts and “understand” to a certain level how words are related to each other, what a certain meaning of a word is and how sentences are constructed.

In this thesis, NLP is used to collect structured data that are included in the radiological report. An oncological classification can then be determined by analyzing this data. This facilitates the check if all necessary information is present in the report, also making it is possible to determine part of the TumorNodeMetastasis (TNM) stage. This application of NLP has been used to initially analyze reports of pulmonary CT scans at the T (tumor) stage only. Subsequently, this algorithm was translated into English so that it is also possible to analyze reports in English. Subsequently, the Dutch algorithm

was also expanded to determine the N (lymph node) stage in both CT reports and PET-CT reports.

The beauty of this research is that it shows that AI can be applied to search radiological records and thus offers the opportunity to improve them. Of course, there are still hurdles to be taken before this can be used flawlessly in practice, but the relatively easy training of currently small amounts of data in combination with the use of rules certainly seems promising. The use of a graphical user interface (GUI) also helps implementation in daily practice. This facilitates analysis of texts, also presenting this analysis and its outcome to the radiologist in a readable and reusable way.

For the future, finding more generic building blocks with NLP through Machine Learning is the next goal. Training textual passages on uncertainties, dimensions, correlations and recognizing important sections are common things that are important for many applications. The valorization of algorithms as described above in external institutions is also important, as there are certainly differences in reporting, vocabulary used, but possibly also in personal choices. The GUI in particular, as well as this valorization, will have to ensure that this NLP tooling can be embedded in clinical practice.

ADDENDUM

Nederlandstalige samenvatting

Dit proefschrift getiteld “The Radiological Report: *A compromise between Structured Reporting and Natural Language Processing*”, beschrijft in twee delen de zoektocht naar hoe het radiologisch verslag en het proces van radiologische verslaglegging verbeterd kan worden.

DEEL 1: Gestructureerde verslaglegging in de radiologie

In het eerste deel wordt de toepassing van gestructureerde verslaglegging besproken. Hoewel inzet gestimuleerd wordt door internationale radiologische organisaties is onvoldoende onderzocht of gestructureerde verslaglegging wel zorgt voor verbetering van het radiologische verslag. Dit komt met name doordat standaardisatie samengenomen met gestructureerde verslaglegging. In vele onderzoeken wordt standaardisatie afgedwongen door middel van gestructureerde verslaglegging door de verslaglegger verplicht invulvelden te laten invullen of te laten kiezen uit een specifiek (drop-down) menu. Echter, het lijkt dat met name deze standaardisatie zorgt voor verbetering van (de inhoud) van de verslagen en dat de invloed van gestructureerde verslaglegging nog twijfelachtig is. Met name ook het gebrek aan onderzoek naar de verschillende vormen van gestructureerde verslaglegging zorgt ervoor dat er nog veel werk te doen is om hier conclusies uit te trekken. Wat wel lijkt te werken is het dwingende karakter van doorvoeren van standaardisatie. Echter, moet dit met gestructureerde verslaglegging?

DEEL 2: Natural Language Processing

In dit tweede deel wordt onderzocht hoe NLP als tegenhanger van gestructureerde verslaglegging ingezet kan worden in de zoektocht naar verbetering van het radiologische verslag. NLP is een vorm van artificial intelligence (AI) waarmee teksten doorzocht kunnen worden, waarbij NLP tot op een bepaald niveau kan “begrijpen” hoe

woorden met elkaar samenhangen, wat een bepaalde betekenis van een woord is en hoe zinnen opgebouwd zijn.

In dit proefschrift is NLP gebruikt om gestructureerd data te verzamelen welke in het radiologische verslag staan. Vervolgens kan door deze data te analyseren een oncologische classificatie bepaald worden. Hierdoor is het mogelijk om te checken of alle nodige informatie aanwezig is in het verslag, en in aanvulling hierop is het mogelijk om een deel van het TumorNodeMetastasis (TNM) stadium te achterhalen. Deze toepassing van NLP is gebruikt om in eerste instantie verslagen van CT long scans te analyseren op enkel het T (tumor) stadium. Vervolgens is dit algoritme vertaald in het Engels zodat het ook mogelijk is om Engelstalige verslagen te analyseren. Daarna is het Nederlandstalige algoritme ook uitgebreid om zo ook het N (lymfeklier) stadium te bepalen in zowel CT verslagen, maar ook PET-CT verslagen.

Het mooie van dit onderzoek is dat het laat zien dat AI toegepast kan worden om radiologische verslagen te doorzoeken en zo de mogelijkheid biedt om deze te verbeteren. Natuurlijk zijn er nog hordes te nemen voordat dit feilloos in de praktijk ingezet kan worden, maar met name het relatief makkelijk trainen van op dit moment kleine hoeveelheden data in combinatie met het gebruiken van regels lijkt zeker veelbelovend. Ook de inzet van een *graphical user interface* (GUI) helpt in het beter kunnen implementeren in de dagelijkse praktijk. Hierdoor is het namelijk mogelijk om teksten te analyseren en deze analyse met uitkomst te laten zien in een leesbare en herbruikbare lay-out voor de radioloog.

Waar we in de toekomst zeker naar willen kijken of we meer generieke bouwstenen kunnen vinden met NLP door middel van Machine Learning. Het trainen van tekstuele passages over onzekerheden, afmetingen, correlaties en het herkennen van belangrijke secties zijn algemene dingen die voor vele toepassingen van belang zijn. Ook het valoriseren van algoritmes als hierboven omschreven in externe instellingen is van belang, aangezien er zeker verschillen bestaan in verslaglegging, gebruikte vocabulaire, maar mogelijk ook van persoonlijke keuzes. Met name de GUI alsook deze valorisatie zullen ervoor moeten zorgen dat deze NLP-tooling beter ingebed kunnen worden in de klinische praktijk.

Impact

Research

The radiological report is the main and most important output of the radiologist as it states the outcome of the performed examination in concordance to the condition of the patient and the suggested diagnosis. However, due to differences in reporting and report content, the value of the radiological report is not always the same. In addition, the reporting process is still the same as in the earliest days, despite suggested reporting improvements like structured reporting (SR) and standardization. Especially these two possible improvements are widely promoted by large radiological societies to increase the value of the radiological report.

As the radiological report is very important, it is necessary to know how the reporting process can be enhanced and why for instance SR is still not implemented in the full field of radiology. This might be caused by difficulties in implementation or full suitability in the whole field and perhaps a negative sentiment among the reporters.

In addition, we need to search for different solutions to improve radiological reporting as the quality of the radiological report is still not improved and SR might not the way to go. Natural Language Processing (NLP), by using Artificial Intelligence (AI), can also extract and analyze free text and might be a substitute for SR in order to improve report quality.

Relevance

This thesis highlights the ongoing search towards improving radiology reporting focusing on structured reporting and the use of NLP. In the first part of this thesis, it is shown that due to difficulties in definitions of SR and its interpretations, a bonanza of scientific papers appeared. Proper setting the definitions for structured reporting and standardized reporting (Chapter 2), will increase its understanding and will allow for more evidence-based research. This is especially important as the current research performed with structured reporting has a low evidence level (Chapter 3). The output

of most studies is beneficial for structured reporting, but this is mostly due to better implementation of standardized reporting and not due to the fact that structured reporting is implemented. These outcomes are important for future implementations and question the promotion of structured reporting by radiological societies, as high-level evidence-based research is still lacking.

NLP has been used in healthcare to structure free text data. As it can also structure free text radiological reports it can be a substitute for structured reporting in radiology. This thesis provides the evidence that in a pilot setting it is possible to extract the Tumor and Nodal (TN-stage) necessary for oncological staging of pulmonary carcinoma out of free text radiological (PET) CT reports according to the TumorNodeMetastasis (TNM) classification system (Chapters 4-6). In addition, and because NLP structures free text data, it is possible to use the separate data components in for instance the radiological report. When combining both applications of NLP it is possible to use the free text data in a structured format as well as add value to the report by adding the described oncological stage. Hereby, a graphical user interface (GUI) is suggested to be a vehicle to improve NLP-processes, as it can highlight NLP results in the reporting process (Chapter 5).

This research does also show that for implementing AI in radiological reporting not always a large amount of data is necessary, but that smaller data sets can suffice, especially when using a rule-based approach. A different advantage of this rule-based approach is that the workflow is known and can be adjusted easily. This is also true for changing the language of the algorithm as is shown in chapter 5. Because it is rule-based only the language needs to be adjusted instead of training the algorithm and its rules again in a different language. Especially this will make the rule-based approach interesting for future research.

Target population

First of all, this thesis is a message to the structured reporting community that the evidence for structured reporting is questionable and that they should look closer to the

reporting process in order to assess its different components. It is advisable to review the efforts done in concordance with the new definitions for structured reporting and standardization, as standardization of the report content alone seems sufficient. For future research it is important to implement studies to investigate in which cases the structured reporting format is beneficial and in which it is not.

Secondly this thesis is interesting for the radiology reporter as it is important to get more insights in the reporting process as well as getting more insights in the problem of inconsistent reporting. After all, there is still much to gain in enhancing the value of the radiological report, and perhaps the possible solutions are much easier than letting AI and PACS vendors do their magic. Nevertheless, PACS-vendors and NLP experts should cooperate in searching evidence, solutions and applications for free text mining. After all, it will be a real improvement in radiology reporting when structuring data and adding specific information is integrated in the PACS system. Especially we postulate that a GUI will increase the acceptability and understanding NLP dependent tools among radiology reporters.

Finally, if we look beyond radiology and where NLP tools can be implemented in daily clinical routine, it might assist with all kinds of administrative tasks that current healthcare is facing. NLP solutions than can be a problem solver to overcome administrative burdens and thereby allowing healthcare employees to use more time for patient care.

Future

The overall aim should be to combine the image information of the radiological examination into the radiological report and use this combined information to improve the final radiological report. The GUI should be the central point of interaction and text as well as image algorithms should assist to increase the value of the radiological report. This is something we are aiming for, probably since the discovery of the X-ray by Wilhelm Conrad Röntgen.

Curriculum Vitae

Martijn Nobel was born in Dodewaard on 25 January 1987. After finalizing his high school at the Stedelijk Gymnasium in Nijmegen, he started in 2005 with Pharmaceutical Sciences at the University of Utrecht. In 2006 he switched studies and started Medicine at the University of Utrecht. In 2013, he commenced his residency Radiology at RadboudUMC, supervised by prof. dr. Schultze Kool and dr. Peters-Bax. In 2014 he moved to the Maastricht UMC+ to finish his residency, supervised by prof. dr. De Haan and dr. Jacobi-Postma. His scientific career officially began in 2015 with a part-time PhD trajectory on structured reporting and AI in radiology at the School of Health Professions Education of Maastricht University under supervision of prof. dr. Robben and prof. dr. Dekker. After finishing residency in Radiology in 2018, he continued his career with a fellowship Neuroradiology/Head and Neck as well as a fellowship Artificial Intelligence at the Maastricht UMC+. In 2019, he visited to collaborate with the Artificial Intelligence in Medicine (AIM) team of prof. dr. Aerts at Harvard-Massachusetts General Hospital. Since 2020 he is working as an attending radiologist at the Maastricht UMC+, focusing on head and neck, neuroradiology and forensic radiology. Besides his clinical work, he is also involved in hospital-wide initiatives on AI and standardizing and structuring the (research) infrastructure at the Maastricht UMC+ in his function as Medical Information Officer of the department of Radiology and Nuclear Medicine. Since 2023, he is also managing the division of Forensic Radiology at the Maastricht UMC+.



List of Publications

Nobel JM, Kok EM, Robben SGF. Redefining the structure of structured reporting in radiology. *Insights Imaging*. 2020;11(1):10. doi: 10.1186/s13244-019-0831-6.

Nobel JM, Puts S, Bakers FCH, Robben SGF, Dekker ALAJ. Natural language processing in Dutch free text radiology reports: challenges in a small language area staging pulmonary oncology. *J Digit Imaging*. 2020;33(4):1002-1008. doi: 10.1007/s10278-020-00327-z.

Nobel JM, van Geel K, Robben SGF. Structured reporting in radiology: a systematic review to explore its potential. *Eur Radiol*. 2022;32(4):2837-2854. doi: 10.1007/s00330-021-08327-5.

Nobel JM, Puts S, Weiss J, Aerts HJWL, Mak RH, Robben SGF, Dekker ALAJ. T-staging pulmonary oncology from radiological reports using natural language processing: translating into a multi-language setting. *Insights Imaging*. 2021;12(1):77. doi: 10.1186/s13244-021-01018-1.

Puts S, **Nobel JM**, Zegers C, Bermejo I, Robben SGF, Dekker ALAJ. How natural language processing can aid with pulmonary oncology Tumor Node Metastasis staging from free-text radiology reports: algorithm development and validation. *JMIR Form Res*. 2023;7:e38125. doi: 10.2196/38125.

Nobel JM, Puts S, Krdzalic J, Robben SGF, Dekker ALAJ. Natural language processing algorithm used for staging pulmonary oncology from free-text radiological reports: including PET-CT and validation towards clinical use. *Submitted*.

Repository of the rule based NLP TN-stage algorithm
<https://gitlab.com/medstruct>

Other

Gietema HA, Zelis N, **Nobel JM**, Lambriks LJG, van Alphen LB, Oude Lashof AML, Wildberger JE, Nelissen IC, Stassen PM. CT in relation to RT-PCR in diagnosing COVID-19 in The Netherlands: A prospective study. *PLoS One* 2020;15(7):e0235844. doi: 10.1371/journal.pone.0235844.

de Nooijer RA, **Nobel JM**, Arets HG, Bot AG, van Berkhout FT, de Rijke YB, de Jonge HR, Bronsveld I. Assessment of CFTR function in homozygous R117H-7T subjects. *J Cyst Fibros* 2011;10(5):326-32. doi: 10.1016/j.jcf.2011.03.009.

Dankwoord

Bij dezen wil ik graag iedereen hartelijk danken die mij hebben geholpen bij het tot stand komen van dit proefschrift. Ik denk niet dat ik volledig zal zijn in mijn bedankjes, maar mocht ik iemand vergeten zijn dan is dat zeker geen onwil.

Prof. dr. S.G.F. Robben, geachte promotor, Simon, hartelijk dank voor de begeleiding in de afgelopen jaren. Waar je in 2015 begon als mijn promotor ben je nu naast mijn promotor, ook mijn mentor, kamergenoot, maar vooral iemand waarmee ik goed kan sparren over allerlei serieuze, maar vooral ook niet serieuze zaken. Het is altijd fijn om een luisterend oor te hebben om zaken te bespreken en deze kritisch tegen het licht te kunnen houden, om vervolgens weer lekker te bagatelliseren of uit z'n verband te trekken met een goede dosis sarcasme en/of cynisme. Op dit niveau functioneer ik goed. Ik hoop dat we nog even zo door kunnen gaan!

Prof. dr. A.L.A.J. Dekker, geachte promotor, beste André, ook jou wil ik hartelijk danken voor de begeleiding de afgelopen jaren. Naast het feit dat je mijn promotor bent, zien we elkaar ook geregeld in andere setting binnen de muren van het Maastricht UMC+. Hierbij ontstaan vaak mooie ideeën, welke vaker niet dan wel (direct) uitgevoerd worden. Echter, aan innovatie en het toepasbaar maken van AI in de gezondheidszorg valt nog wel wat te verbeteren, waardoor we elkaar ook na mijn promotie geregeld tegen zullen komen. Met name gedurende de eerste COVID-golf in 2020 hebben we nauw samengewerkt om een AI algoritme naar Maastricht te krijgen. Dit was een erg leuke en inspirerende periode (relatief dan natuurlijk), waarin ik je heb leren kennen als een fijn en laagdrempelig te benaderen persoon. Proost op je muffins en chihuahua's!

Sander, meneer Puts, mede-auteur in alle NLP-artikelen en natuurlijk mijn technische steunpilaar. Wij vormen sinds het begin van onze samenwerking een mooi blok van enerzijds klinische ideeën en anderzijds jouw technische kennis om van een idee een werkende toepassing te maken. Ik kan mij de periode in Boston goed herinneren, waarbij ik vroeg in de morgen opstond en jij tot diep in de nacht opleef om samen aan

ons Engelse algoritme te werken. Een tijd waarin we goed de flow te pakken hadden. Ook nu is het nog altijd fijn samenwerken en we gaan zeker samen nog even door, dank!

Prof. dr. J.E. Wildberger en drs. F. Bakers, als afdelingshoofd en toenmalig plaatsvervangend afdelingshoofd wil ik ook jullie danken. Joachim, jij hebt het uiteindelijk mogelijk gemaakt dat ik als AIOS heb kunnen starten aan dit promotietraject. Daarna heb ik een gecombineerd Neuro/Hoofd-hals & AI fellowship kunnen doorlopen binnen de afdeling Beeldvorming en we meer gefocust hebben op de implementatie van AI in de klinische praktijk. Frans, hier kwam jij ook in beeld gezien jouw uitgebreide ICT-werkzaamheden waarbij ik “ons” visiedocument AI mede zie als startpunt voor de verschillende processen welke binnen de afdeling Beeldvorming, maar ook ziekenhuis-breed in gang gezet zijn. Ik wil jullie beiden danken en hoop voor de toekomst, met de veranderende takenpakketten van eenieder, dat we nog mooie stappen kunnen maken!

Prof. dr. H.J.W.L. Aerts, dr. R. Zeleznik and J. Weiss. During my stay in Boston at Harvard you facilitated me throughout the NLP research. It is nice to have been working together also having shared some less serious time. Hugo, thanks for giving me the opportunity to come over. I hope we can do more research when we look at clinical implementation of different AI tools. Roman, nice to see that you are post-doc now. Nevertheless, I hope you have new shoes by now. Jakob, great you are working in the field of interventional radiology, and good luck with finishing your PhD!

Koos van Geel, ook jou wil ik kort danken (en ik weet dat jij dit op waarde kunt schatten), dank! Ellen Kok, hoewel wij enkel in het begin hebben samengewerkt, denk ik daar met plezier op terug. Met name je grondige feedback is iets wat ik me goed kan herinneren!

Ook mijn opleiders in het Maastricht UMC+, dr. A.A. Jacobi-Postma en prof. dr. M.W. de Haan wil ik danken. Allereerst omdat ik naar Maastricht kon komen, maar daar ben ik ook dank voor verschuldigd richting het RadboudUMC, en dat ik vervolgens ook

parttime heb kunnen beginnen aan mijn onderzoek wat geleid heeft tot dit boekje. Dank voor de opleiding, check!

Collega's neuroradiologie en/of forensische radiologie. Ook jullie dank voor het begrip als ik weer eens een afspraak of andere verplichting had waar ik heen moest. Ook jullie hebben aan de totstandkoming van dit boekje meegewerkt – dank hiervoor.

Ook wil ik graag de dames van het secretariaat Beeldvorming bedanken. Nicole, Jolanda, Christianne en Elfie, dank voor jullie steun de afgelopen jaren en een praatje op z'n tijd. Ik denk dat ik jullie ook de komende tijd nog niet kan missen...

Bas, Juul, Robbert en Wouter, vrienden vanuit onze studietijd in Utrecht. Hoewel we ondertussen niet meer zo bij elkaar om de hoek wonen, weten we elkaar toch met enige regelmaat alsnog te vinden. Hoewel we dit niet van tevoren voorzien hadden, lijkt radiologie toch een verbindende factor te zijn binnen onze groep! Altijd leuk om jullie weer te zien of te spreken. Pizza eten blijft toch wel echt onze specialiteit.

Wijnand, oud-huisgenoot, buur! Jaren hebben we aan de Leeuwerikstraat in Utrecht gewoond en samen gekookt, geklust, geschaatst en vele andere dingen gedaan. Ik waardeer je optimisme en je altijd goede humeur. Ook al wonen we wat verder van elkaar, zo voelt het niet.

Pa en ma, natuurlijk wil ik jullie ook heel, heel erg bedanken. Hoe jullie mij altijd gestimuleerd hebben en geholpen in mijn keuzes, dat is fenomenaal! Het maakt niet uit waarmee, maar jullie staan altijd voor mij (intussen ons) klaar met objectief advies of simpelweg voor een dagje oppas. Hopelijk kunnen jullie lekker gaan genieten van jullie pensioen en kunnen jullie nog maar vaak langskomen in het Zuiden!

Marieke, hoewel wij toch heel iets anders doen is het ook leuk om te zien dat jij zo goed bezig bent. Mooi om te zien dat je in je managementfunctie zoveel kunt betekenen voor je schoolkinderen. Ook altijd leuk om te zien dat onze kinderen het zo goed met elkaar kunnen vinden, we hebben al heel wat leuke dagen gehad met z'n allen aan het strand!

Floris en Louise, hoewel jullie zeker niet op elkaar lijken, lukt het jullie aardig om mij scherp te houden en op allerlei vlakken uit te dagen. Het is altijd fijn om jullie om me heen te hebben en jullie te zien genieten van de kleine dingen, zoals samen een pannenkoekje bakken of met de Lego bezig zijn. Sinds kort hebben we de nieuwe activiteit om samen ons rondje Bemelerberg te lopen. Top om te zien dat jullie ook zo van de natuur houden en, hoewel jullie het waarschijnlijk niet (te hard) zullen merken of willen weten, ben ik best wel trots op jullie kleine mensjes!

Anna-Jasmijn, Apples, lieve vrouw. Sinds wij elkaar kennen hebben we al zoveel dingen meegemaakt tijdens onze studententijd, reizen samen en tijdens onze tijd in opleiding. Hoe mooi deze ervaringen ook waren en zijn, toen we Floris en later Louise kregen, heeft dit weer een supermooie wending gegeven aan ons leven. Zeker ook jij hebt mij geholpen bij het promoveren, het redigeren en het doorlezen van met name alle versies van mijn stukken. Duizendmaal dank hiervoor! Een hoogtepunt van mijn promotietraject was toch wel ons bezoek aan Boston en hoe fijn was dat om met het gehele gezin te ondernemen. Ik hoop dat we met z'n allen nog vele mooie herinneringen mogen maken in Bemelen en all over the world!

