Primary care datasets for early lung cancer detection: an AI led approach

Goce Ristanoski[1], Jon Emery[2,3], Javiera Martinez Gutierrez[2,4], Damien McCarthy[2] and Uwe Aickelin[1]

[1] School of Computing and Information Systems, The University of Melbourne, Australia
[2] Department of General Practice and Centre for Cancer Research, Medicine, Dentistry and Health Sciences, The University of Melbourne, Australia
[3] Victorian Comprehensive Cancer Centre, Melbourne, Australia
[4] Department of Family Medicine, School of Medicine. Pontificia Universidad Católica de Chile
`gri@unimelb.edu.au`

Abstract. Cancer is one of the most common and serious medical conditions, with significant challenges in the detection of cancer originating from the non-specific nature of symptoms and very low prevalence. For general practitioners (GPs), this can be particularly important, as they are the primary contact for patients for most medical conditions. This places high significance on using the data available to a GP to design decision support tools that will aid GPs in detecting cancer as early as possible. With pathology data being one of the datasets available in the GP electronic medical record (EMR), our work targets this type of data in an attempt to incorporate an early cancer detection tool in existing GP practices. We focus on utilizing full blood count pathology results to design features that can be used in an early cancer detection model 3 to 6 months ahead of standard diagnosis. This research focuses initially on lung cancer but can be extended to other types of cancer. Additional challenges are present in this type of data due to the irregular and infrequent nature of doing pathology tests, which are also considered in designing the AI solution. Our findings demonstrate that hematological measures from pathology data are a suitable choice for a cancer detection tool that can deliver early cancer diagnosis up to 6 months ahead for up to 8 out of 10 patients, in a way that is easily incorporated in current GP practice.

Keywords: Early lung cancer detection, Primary care data, Explainable AI

1 Introduction

Through one's medical history we come in contact with our General Practitioners (GPs) far more often than we do with other medical staff, particularly specialists. The resources and technology available at the GP practices are, however, more limited to those in hospital specialist care. GPs play a key role in diagnosis of serious diseases, but this can be challenging due to the fact that symptoms alone are poorly predictive, especially for uncommon conditions in primary care. Decision support tools could potentially contribute to flagging patients at increased risk of serious disease and prompting further referral or investigation.

One of the medical conditions that can have serious consequences on patient's lives depending on the time of diagnosis is cancer. In Australia, there are more than 144 000 cancer patients who were diagnosed in 2019 alone. Early detection of cancer by GPs is challenging if symptoms alone are used and patients existing history is underutilized. In the last 15 years, the research in the epidemiology of cancer symptoms in primary care data has grown, with many findings demonstrating how advanced analysis and combinations of different symptoms and tests from a patient's medical history can be used to assess cancer risk [1,2,3]. If patients have a regular GP they visit, having just 2-years' worth of patient data can be sufficient in some cases to combine several different tests into a risk prediction model that can provide the initial diagnosis around 3 months ahead of the current practice [3,4]. This may not seem like too long a period at first, but with studies showing how every additional month of an undiagnosed cancer can increase the mortality rate for certain types of cancer [5], establishing early diagnosis at the GP's office is even more important.

Pathology results are one of the most common types of data that exist in the patient EMR that is readily available to a GP. This opens the opportunity to investigate if some of the blood tests can be associated with certain types of cancer. Recent research highlights raised platelet count (thrombocytosis) as a predictor of cancer risk [6,7], but there have been no specific studies that focus on understanding how to introduce a more advanced AI component into cancer detection and how it can be adapted to current pathology data in the GP's EMR. Our work places a strong emphasis on this, allowing for both interpretability of our results and easy application and usage.

The full blood count test results we investigate as a potential input for a Machine Learning/AI model are: Platelet count, MCV (Mean Corpuscular Volume), MCH (Mean Corpuscular Hemoglobin - average mass of hemoglobin per red blood cell), MCHC (Mean Corpuscular Hemoglobin Concentration - concentration of hemoglobin in a given volume of packed red blood cell) and RDW (Red blood cell distribution width). Platelet count is already associated to lung cancer from other studies, and this set of features allows to develop an initial approach to use pathology results in cancer detection, providing opportunity to expand the list of pathology test metrics with more metrics in future work. We focused on lung cancer patients' pathology results as this is a common cancer and patients often have multiple pathology results; lung cancer has a high mortality rate and could benefit from an early cancer detection model.

The work we conducted places a heavy focus on delivering an initial cancer diagnosis early as possible, which is why we developed our models to make predictions 3-months and 6-months ahead of current practices. We attempted to design a model that could flag a diagnosis 6 months in advance specifically to aid in early detection of cancer for high risk patients - patients who did not survive the cancer, who potentially have most to gain from earlier detection.

We present our work with the following contributions:

- We discuss the ideas of using pathology results in an AI model for cancer detection and show reasoning behind this hypothesis.

- We demonstrate how the metrics listed above can be relevant to cancer patients.
- We address the type and structure of pathology results data available to a GP to design features that are easy to implement and use in a cancer detection model.
- We present AI models with performance that shows promising results in detecting cancer, especially for high risk patients
- We list future opportunities that can improve this type of work even further with little modification and wide application.

2 Related work

With more than 2.9 million deaths worldwide associated to lung cancer in 2018, it has become imperative to find additional ways to better detect the early symptoms of lung cancer and provide timely diagnosis [8]. The main challenge about the symptoms however is that they can vary from one patient to another and can take even up to 2 years for the symptoms to be visible enough to have them attributed to cancer [9]. Raised platelet count, or thrombocytosis, has been shown to be an indicator for cancer, with differences in the results for biological male vs female patients – male cancer patients were 50% more likely to have thrombocytosis than female patients [6,10]. This resulted in the practice of referring patients with thrombocytosis to an x-ray scan in an attempt to detect the cancer patients promptly [11]. Anemia has also been shown to have association with lung cancer, with slightly higher presence in male patients as well [2,10], which brings us to our hypothesis of investigating blood count results in combined scenario.

Designing risk prediction models with individual metrics have been investigated to a good extent [12,13,14], but without strong emphasis on combining several metrics into single model, or considering the application range of the prediction models in GP offices. Reviews on the use of primary care data for cancer prediction with other types of primary care datasets are also indicating that blood results are increasingly popular for the task [15,16], and with indications that lung cancer patients tend to have blood tests more often [17], it provides fertile grounds for introducing AI models in the bigger picture.

3 Dataset description

3.1 NPS MedicineInsight

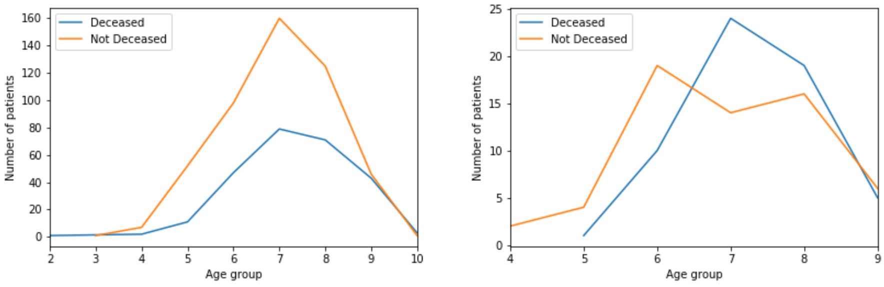
The Australian Government Department of Health (DoH) established the NPS MedicineInsight initiative as a nationally representative primary care dataset that can be used by academic researchers in attempt to deliver new research findings that can improve medical practices. The NPS MedicineInsight contains patients records from more than 500 general practices and 5000 GP providers, which includes more than 8 million recorded diagnoses, 23 million prescriptions, 32 million encounters and 85 million pathology test results [18]. For our research work, we obtained the lung cancer patients cohort, as well as a non-cancerous patients’ cohort as a control group.

With an extensive amount of records and results from pathology tests, we focused our work on the five blood test metrics listed earlier: Platelet count, MCHC, MCV, MCH and RDW. We looked at the out of range records for these metrics, with the standard range being: platelet count of 150-450 x 10^9/L, MCV of 80-98 fL, MCH of 28-32 pg/cell, MCHC of 330-370 g/L, RDW of 12.2%-16.1% F/ 11.8%-14.5% M. Our models used an out of range value in at least one of these metrics as a trigger for classification, meaning that lung cancer patients that have no out of range values unfortunately were not assessed for early detection. The analysis showed that around 20% of the patients per each metric had a record with an out of range value for that metric, so combining several metrics increased the total subset of lung cancer patients suitable for early detection. Subsequently, we only considered non-cancerous patients with out of range values as a control group.

3.2 Cancer patient’s analysis

The available lung cancer patient cohort showed a very interesting pattern compared to other patients. One of the things we noticed initially was that not only did lung cancer patients had around 20% out of range value for a given blood test metric, they also had on average 3 times more tests taken in the two year period before cancer diagnosis than patients that had no out of range results, allowing both better quality in data and initial indication of use of pathology results for early diagnosis.

Another interesting aspect of our analysis showed that not only there were out of range tests for a good portion of cancer patients, but that the mortality rate for patients with out of range tests was much higher than for patients with no out of range results. Shown in **Fig. 1** is an example for patients with out of range results for RDW compared to patients with no out of range results and the mortality figures per age group (group 2=20-29 y.o. etc). We can observe higher mortality ratio for patients with out of range, showing that even if we can only include patients with out of range results in the final model, these patients are high risk patients and they may benefit from early cancer detection the most.



a) b)

Fig. 1. Patients that survive vs. patients that did not survive cancer based on out of range results for RDW: a) Patients with no out of range results; b) Patients with out of range results 3 months prior diagnosis date

4 Features design and methods selection

4.1 Uncertainty based features design

Pathology tests are readily available to GP to order, but as we could see in the previous analysis, the number of tests per patient can vary – lung cancer patients that had out of range results had more tests than other patients which worked in favor of our work to some extent. The frequency and regularity of these tests is not a matter of standardized practices: GPs issue a request for test when patients visit them, and this is something that is irregular and driven by a range of factors. This poses some limitations on the quality of data as well as the use of this type of data for AI models.

Our approach was to handle this uncertainty by using time periods and occurrences of out of range tests results to incorporate some structure in the features and allow for use of pathology data without any special need for its format other than the current ones used in practice. For the lung cancer patients, from the initial diagnosis date recorded at the GP clinic, we took the pathology tests within the two-year period prior to that date. We then represented the occurrence of any out of range results for each of the five listed metrics in the periods of 24-18, 18-12, 12-6 and 6-3 months before the cancer diagnosis date. The occurrence of each metric per individual period formed one original feature, with 0 meaning no occurrence of out of range result for that metric for the given time period, and 1 meaning at least one occurrence of out of range result for that metric in that time period. This created data suitable for 3 months before diagnosis, and by removing the features with the 6-3 months period we could also perform a 6 months before diagnosis feature. For the control group of non-cancerous patients we selected the period of 2016-2017 and the same features were calculated for that period. We did not consider multiple occurrences within one time period as often pathology tests can be issued subsequently, and this would bring no new information to our models.

4.2 Soft out of range results

The normal ranges for each of the hematological measures were listed earlier, and based on some of the test results, we included the results at the very end of the normal ranges as soft out of range. For example, platelet count is most commonly defined within normal range of 150-450 thousand platelets per microliter of blood, so patients with results of 150 will be within the range, but patients with results 149 are out of range. In order to allow patients with results of 150 or just above it to still be considered as out of range we defined soft range as being the 2.5% ends from within the lowest and highest values. Using the platelets example, 2.5% of the 450-150=300 is 7.5 units, so the soft ranges would be (157.5-442.5).

This means that we ended up with more test results being out of range and potentially more patients being suitable to be considered for early diagnosis. For the 3 months before diagnosis, using the standard range we had 592 patients within our dataset, and that number increased to 683 with the soft out of range definition.

4.3 Additional features

Besides the original features for the five metrics for each time period listed above, we combined some more features based on those and other patient data to allow some more temporal and quantitative aspects to be included in the algorithm. These were:

- Summary of occurrences per blood test metric
- Summary of occurrences of any metrics over a 3- or 6-month pre-diagnosis period
- Separating the out of range values into two separate features for upper and lower threshold out of range
- Separating the previous features per biological sex
- Separating the previous features per age group

By using this feature set, we could get a clearer view of the importance of occurrences vs. frequency of out of range results, both total and per individual age group or biological sex.

5 Model selection, experiments and results

5.1 Model selection

The use of pathology data in the features listed above not only handled the uncertainty in the data that originated from the irregularity of pathology tests, but it also provided another crucial contribution: it allowed us to see if the individual original features or the combinatory ones had more useful information. In order to allow even more interpretability in both the final performance and the relevance of the features, we used decision tree style models: Decision Tree, AdaBoost, LightGBM and XGBoost. We also used an ensemble approach to check for additional performance evaluation: a stack model that uses the forecasts of the other classifiers as an input, as well a simple ensemble with the OR logic between all the classifiers.

We were interested to see how our models performed in correctly classifying the lung cancer patients. We wanted to achieve both high values for True Positive Rate (TPR) and True Negative Rate (TNR), and also from all the predicted positives we wanted the cancer patients to have the highest portion (Positive Predictive Value, PPV). We had a total of 592 patients for the 3 months ahead early diagnosis, and 683 patients for the same diagnosis with soft out of range features. For the 6 months ahead early diagnosis, we had 499 patients total for both standard range and soft out of range.

The use of different ratios of non-cancerous patients:cancer patients allowed us to see how the TPR and TNR changed and if we could avoid having lots of false positives. We used the ratio of 1:1, 1.5:1 and 2:1, and suggested not going higher than 4:1 in order to avoid issues due to an imbalanced dataset. The chi-squared statistic for ranking the top features was used, and we showed the average performance when using 41-54 features.

Our cancer patients were all 50+ years old, and we also added a subset the 50-79 years range to allow for better quality of data as patients aged 80+ had different frequencies of pathology tests and could have more health issues that made it difficult to differentiate between cancer based out of range pathology tests and other conditions out of range tests.

5.2 Experiments and results

The results presented in **Table 1** and **Table 2** show the average performance of the models with over 14 runs, with 41-54 features used per run. The standard deviation over each metric was rarely higher than 0.01, so the performance was quite consistent per each run. We investigated the impact of three ratios of non-cancerous patients to cancer patients (1:1, 1.5:1 and 2:1) and in all cases the ratio of 1:1 showed best results for TPR and only those figures are presented here. As we increased the ratio, the value of TPR dropped in all cases while TNR (True Negative Rate) increased to values of 0.9. The PPV value also shows good performance, and it is closely matched with the Negative Predictive Value (NPV).

The results confirm that not only we were able to provide 3 months ahead forecast with our pathology results with good accuracy (7 to 8 out of 10 correctly classified patients) but we could also provide similar accuracy with the 6 months forecast as well. The individual models when combined in an Ensemble with OR logic (if one model classifies a 1, the ensemble outputs 1) performed well for the 3 months ahead forecast, but not as good for the 6 months ahead. The Stack model did not seem to suffer this issue. Still, a 7 out of 10 forecast delivered 6 months ahead with only 5 metrics is very promising in the future use of the pathology results in primary care data for early cancer detection.

The performance of the 6 months ahead forecast was also satisfactory in the prediction of the high-risk patients task. We can observe from **Fig. 2** that for both regular out of range and soft out of range, the percentage of deceased patients in the correctly classified cancer patients was higher than the percentage of deceased patients in the false negative forecasts: in some cases nearly 45% of the patients in the correct classifications were high risk patients that were deceased within 4 years of the cancer diagnosis, and this number was as low as 35% in the false negative forecasts. This shows that our models were able to detect the cancer patients that can benefit from an early diagnosis the most.

Table 1. Performance metrics for data with regular range

All samples	**3 months ahead**				**6 months ahead**			
Classifier	**TPR**	**TNR**	**PPV**	**NPV**	**TPR**	**TNR**	**PPV**	**NPV**
AdaBoost	0.686	0.722	0.711	0.697	0.684	0.679	0.680	0.682
DecisionTree	0.613	0.747	0.708	0.659	0.586	0.692	0.656	0.626
Ensemble	0.807	0.619	0.679	0.763	0.784	0.565	0.643	0.723
LightGBM	0.705	0.710	0.708	0.707	0.684	0.686	0.685	0.684
Stack	0.705	0.710	0.708	0.707	0.684	0.686	0.685	0.684
XGB	0.722	0.748	0.742	0.730	0.671	0.715	0.702	0.685
Under80								
Classifier	**TPR**	**TNR**	**PPV**	**NPV**	**TPR**	**TNR**	**PPV**	**NPV**
AdaBoost	0.650	0.690	0.678	0.664	0.617	0.684	0.661	0.642
DecisionTree	0.594	0.730	0.688	0.643	0.541	0.670	0.621	0.594
Ensemble	0.770	0.573	0.643	0.714	0.755	0.569	0.637	0.701
LightGBM	0.664	0.717	0.701	0.681	0.650	0.653	0.652	0.651
Stack	0.665	0.718	0.702	0.682	0.649	0.653	0.652	0.651
XGB	0.655	0.762	0.733	0.689	0.652	0.717	0.697	0.704

Table 2. Performance results for metrics with soft range

All samples	**3 months ahead**				**6 months ahead**			
Classifier	**TPR**	**TNR**	**PPV**	**NPV**	**TPR**	**TNR**	**PPV**	**NPV**
AdaBoost	0.659	0.760	0.733	0.690	0.665	0.712	0.698	0.680
DecisioTree	0.622	0.765	0.725	0.669	0.601	0.704	0.670	0.638
Ensemble	0.783	0.661	0.698	0.753	0.772	0.595	0.656	0.723
LightGBM	0.686	0.758	0.739	0.707	0.665	0.711	0.697	0.680
Stack	0.686	0.758	0.739	0.707	0.665	0.711	0.697	0.680
XGB	0.657	0.775	0.745	0.694	0.639	0.742	0.712	0.673
Under 80								
Classifier	**TPR**	**TNR**	**PPV**	**NPV**	**TPR**	**TNR**	**PPV**	**NPV**
AdaBoost	0.689	0.732	0.720	0.703	0.592	0.717	0.676	0.639
DecisioTree	0.591	0.689	0.656	0.628	0.537	0.646	0.603	0.583
Ensemble	0.781	0.613	0.668	0.737	0.738	0.585	0.640	0.692
LightGBM	0.675	0.680	0.679	0.677	0.620	0.649	0.639	0.631
Stack	0.675	0.680	0.679	0.677	0.620	0.649	0.638	0.631
XGB	0.670	0.741	0.721	0.692	0.628	0.704	0.680	0.655

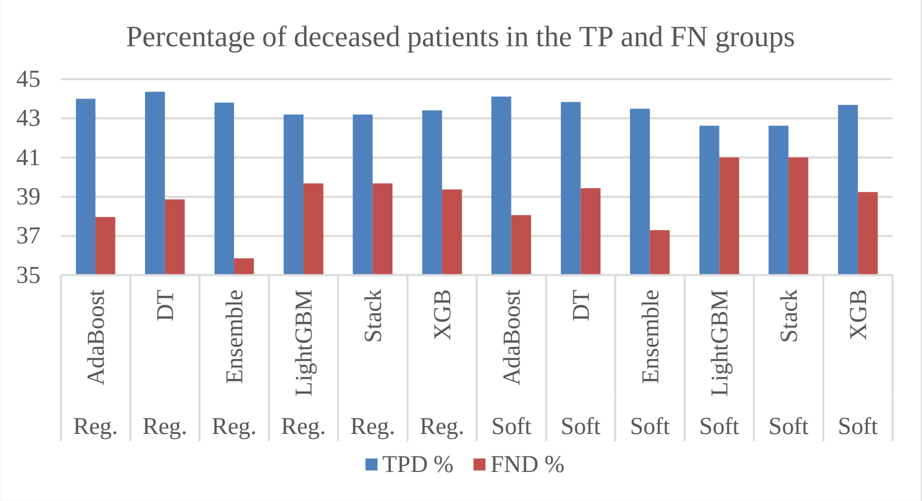


Fig. 2. Percentage of deceased patients in the True Positives (TP) and False Negatives (FN) groups per classification model and out of range type

6 Conclusion

The work presented in this paper demonstrates the opportunities to use currently underutilized set of data for early cancer detection: A primary care dataset containing pathology results. Not only do we justify the reasoning behind the use of full blood test metrics for early cancer detection, but we also handle the challenge of the data containing records at irregular and infrequent time periods. By using features that represent both temporal and quantitative values in the out of range results, we were able to predict lung cancer diagnosis up to 6 months ahead of time, with models that required no modification to current GP practices and would be relatively easy to implement in clinics for both lung cancer detection and other types of cancer as well.

This work opens opportunities for further research in areas such as more high-risk patient focused forecast, inclusion of other pathology tests, and potentially incorporating social and economic features in the AI models. Based on availability of additional data about the stage of cancer at the time of detection and hospital treatment, we may further deliver more insights by using pattern detection and visualization methods to determine the most descriptive features in the pathology tests per different type of cancer or patient cohort.

References

1. Hamilton, W. (2009). The CAPER studies: Five case-control studies aimed at identifying and quantifying the risk of cancer in symptomatic primary care patients. British Journal of Cancer, 101, S80–S86.

2. Hippisley-Cox, J., & Coupland, C. (2011). Identifying patients with suspected lung cancer in primary care: Derivation and validation of an algorithm. British Journal of General Practice, 61(592).
3. Corner J, Hopkinson J, Fitzsimmons D, et al. Is late diagnosis of lung cancer inevitable? Interview study of patients' recollections of symptoms before diagnosis. Thorax 2005;60:314-319.
4. Iyen-Omofoman, Barbara, et al. "Using socio-demographic and early clinical features in general practice to identify people with lung cancer earlier." Thorax 68.5 (2013): 451-459.
5. Hannah TP et al. Mortality due to cancer treatment delay: systematic review and meta-analysis BMJ 371 (2020).
6. Bailey, S. E. R., Ukoumunne, O. C., Shephard, E., & Hamilton, W. (2017). How useful is thrombocytosis in predicting an underlying cancer in primary care? A systematic review. Family Practice. Oxford University Press.
7. Bailey, S. E. R., Ukoumunne, O. C., Shephard, E. A., & Hamilton, W. (2017). Clinical relevance of thrombocytosis in primary care: A prospective cohort study of cancer incidence using English electronic medical records and cancer registry data. British Journal of General Practice, 67(659), e405–e413.
8. Shapley, M., Mansell, G., Jordan, J. L., & Jordan, K. P. (2010, September). Positive predictive values of ≥5% in primary care for cancer: Systematic review. British Journal of General Practice.
9. Bjerager M, Palshof T, Dahl R, et al. Delay in diagnosis of lung cancer in general practice. Br J Gen Pract 2006;56:863–8.
10. World Health Organization. Cancer Fact Sheets. Available at https://www.who.int/newsroom/fact-sheets/detail/cancer
11. Victoria Cancer Council. I-PACED (Implementing Pathways for Cancer Early Diagnosis) resources available at https://www.cancervic.org.au/downloads/resources/factsheets/AL1720_OCP_I-PACED_Lung_FINAL.pdf
12. ten Haaf, Kevin, et al. "Risk prediction models for selection of lung cancer screening candidates: a retrospective validation study." PLoS medicine 14.4 (2017).
13. O'Dowd, Emma L., et al. "What characteristics of primary care and patients are associated with early death in patients with lung cancer in the UK?." Thorax 70.2 (2015): 161-168.
14. Weller, D. P., M. D. Peake, and J. K. Field. "Presentation of lung cancer in primary care." NPJ primary care respiratory medicine 29.1 (2019): 1-5.
15. Goldstein, Benjamin A., et al. "Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review." *Journal of the American Medical Informatics Association* 24.1 (2017): 198-208.
16. Schmidt-Hansen, Mia, et al. "Lung cancer in symptomatic patients presenting in primary care: a systematic review of risk prediction tools." *British Journal of General Practice* 67.659 (2017): e396-e404.
17. Bradley, Stephen H., Martyn PT Kennedy, and Richard D. Neal. "Recognizing lung cancer in primary care." Advances in therapy 36.1 (2019): 19-30.
18. NPS MedicineWise Annual Report 2019-20, resources available at https://www.nps.org.au/about-us/reports-evaluation.

Author/s:
Ristanoski, G;Emery, J;Martinez Gutierrez, J;McCarthy, D;Aickelin, U

Title:
Primary care datasets for early lung cancer detection: an AI led approach

Date:
2021

Citation:
Ristanoski, G., Emery, J., Martinez Gutierrez, J., McCarthy, D. & Aickelin, U. (2021). Primary care datasets for early lung cancer detection: an AI led approach. Tucker, A (Ed.) Abreu, PH (Ed.) Cardoso, J (Ed.) Rodrigues, PP (Ed.) Riano, D (Ed.) Artificial Intelligence in Medicine, 12721 LNAI, pp.83-92. Springer. https://doi.org/10.1007/978-3-030-77211-6_9.

Persistent Link:
http://hdl.handle.net/11343/270178