AI Based Cancer Detection Models Using Primary Care Datasets

Goce Ristanoski[1], Jon Emery[2], Javiera Martinez Gutierrez[2,3], Damien McCarthy[2], and Uwe Aickelin[1]
[1] Department 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] Department of Family Medicine, School of Medicine, Pontifical Catholic University of Chile, Chile
Email: {gri, jon.emery, javiera.martinez, mccarthyd1, uwe.aickelin}@unimelb.edu.au

***Abstract*—Cancer is one of the most common and serious medical conditions with more than 144 000 Australians having been diagnosed with cancer in 2019. The non-specific nature of cancer symptoms and its low prevalence make cancer diagnosis particularly challenging, especially for primary care physicians/General Practitioners (GPs). Ongoing research in cancer diagnosis places a heavy focus on understanding the epidemiology of cancer symptoms. With GPs being the first point of contact for most patients, prediction models using the patient's medical history from primary care data can be a useful decision tool for early cancer detection. Our work both investigates the opportunities to use primary care data, specifically pathology data, for developing such decision tools and tackles the challenges coming from uncertainty in the data such as irregular pathology records. We present opportunities using the results within the frequently ordered full blood count to determine relevance to a future cancer diagnosis. By using several different pathology metrics, we show how we can generate features suitable for AI models that can be used to detect cancer 3 months earlier than current practices. Though the work focuses on patients with lung cancer, the methodology can be adjusted to other types of cancer and other data within the medical records. Our findings demonstrate that even when working with incomplete or obscure patient history, hematological measures contain valuable information that can indicate the potential of cancer diagnosis for up to 8 out of 10 patients. The use of the proposed decision tool presents a way to incorporate pathology data in the current cancer diagnosis practices and to incorporate various pathology tests or other primary care datasets for similar purposes.**

***Index Terms*—explainable AI, early cancer detection, uncertainty in data, feature generation**

I. Introduction

Healthcare systems rely heavily on primary care clinicians to provide preliminary assessments of the health conditions on patients, ensuring both accuracy in diagnosis and optimal referrals to specialist care are maintained. Being the initial point of contact for all medical issues for an entire population, as well as gatekeepers to specialist care is not an easy task. General Practitioners (GPs) commonly use various laboratory tests to help them deliver their diagnosis at the earliest time and ensure patients have the right treatment in the best possible settings. This also means that the recorded tests could deliver a plethora of answers to questions about early detection of a diverse set of medical conditions. The rich medical histories, alongside details about treatment and referrals, create opportunities to use modern AI methods to develop tools that can assist GPs in their work and help many patients get early diagnosis on their conditions and experience better healthcare and health outcomes.

One of the conditions for which early diagnosis by a GP is highly relevant is cancer. It can be very challenging for a GP to provide an early diagnosis for most cancers, as many cancers present with symptoms that have more common benign causes and the symptoms can appear anywhere from 2 years up to several months before the cancer diagnosis [1]. With low prevalence of cancer in primary care, the task of providing an early diagnosis becomes more challenging. Significant amount of research focuses on understanding the epidemiology of cancer symptoms and how it can be used by primary care physicians. Several risk models [2] and risk assessment tools based on combinations of symptoms [3] show promising results, but there is still much work to be done in delivering a standardised decision tool suitable for use over most GP clinics. Some of the approaches tend to investigate more complex patterns of symptoms [4], which can be difficult to accommodate if the data does not have sufficient detail for all patients' medical history.

A full blood count pathology test is one of the most common and standardised types of tests available in primary care, making it a suitable choice for investigating potential relationships between different pathology tests and cancer diagnosis. In our work, we examine the potential to use some of the full blood count tests (MCHC, MCV, MCH, RDW and platelets count) in order to build a decision support tool for early cancer detection. Our work focuses on lung cancer, as it is one of the most common types of cancer which makes it possible to access a reasonable amount of patient data, but the work can be applied to other types of cancer with few modifications. We focus on delivering a cancer diagnosis, 3 months in advance compared with the original time of diagnosis. This is in accordance with some of the other

Manuscript received October 29, 2021; revised February 18, 2022.

doi: 10.12720/jait.13.2.192-197

models available for the same task that also focus on 3-4 months early prediction [4], and furthermore, because even a one month delay in cancer treatment has been proven to increase mortality in several types of cancer, including lung [5]. Nevertheless, the work is suitable for extending the early diagnosis up to 12 months ahead if some of the uncertainties in the data are handled. The findings of this work could be easily implemented in current GP practices which makes the application of the decision support tool of even greater research interest.

The contributions presented in this paper are as follow:

- We present clear indication of an association between the out of range results in several metrics from full blood count tests and patients diagnosed with cancer.
- We show that these metrics, along with some additional demographic features can be associated with early cancer detection.
- We suggest methods to generate relevant features that can be used for early cancer detection models that consider the challenges in the data such as irregular or missing data.
- We deliver satisfactory performance of the trained early cancer detection models and indicate that the models are suitable for widespread use and can be adjusted to other types of cancer.

II. Related Work

Lung cancer is the current leader in cancer related deaths worldwide, with more than 2.9 million cases per year [6]. Early detection of lung cancer is a challenge for most patients, as patients often express mild or unspecific symptoms leading up to two years before the cancer diagnosis is determined [7]. This attributes to a very late diagnosis for most lung cancer patients, making early detection and referral a challenging task for GPs. Research into individual blood test metrics has shown potential to use simpler standardised blood tests as indicators for cancer. Anaemia among patients has demonstrated an adjusted hazard ratio for lung cancer of 1.75 for females and 1.89 for males with anaemia [2], [8]. Raised platelet count (thrombocytosis) has been shown in several studies to be associated with several types of cancer, including lung cancer [9], [10], (23% of males and 14% of females with thrombocytosis in cancer patients compared with 14% and 12% respectively in the general population). In Australia, thrombocytosis has been recently included as a factor for prompt referral to chest x-rays and a lung cancer referral pathway [11].

Several studies have attempted to generate risk prediction tools and algorithms for lung cancer patients using primary care datasets [4], [12], [13]. Other datasets have been used to confirm that cancer patients have more frequent visits to their GP right before the initial diagnosis, including visits with pathology tests [14] This supports our hypothesis of using pathology tests as potential cancer indicators. The use of AI has been investigated using smaller number of blood test metrics as well as specific patients cohorts [15]-[18], so combining several metrics and applying them to a more general patient cohort is the next research challenge that we tackle in this paper.

III. Materials and Methods

A. Dataset Description

The Australian Government Department of Health (DoH) established the NPS MedicineInsight initiative as a nationally representative primary care dataset from more than 500 general practices and 5000+ GP providers. As part of a data program at the Victorian Comprehensive Cancer Centre, we have obtained the Victorian sub-set of NPS MedicineInsight general practice dataset. This dataset includes more than 8 million recorded diagnoses, 23 million prescriptions, 32 million encounters and 85 million pathology test results. Our work focused on patients with lung cancer that during one or several GP visits prior to diagnosis had a pathology test with the following metrics regarding thrombocytosis and anemia: 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).

The hypothesis behind our work is that out of range results in these metrics may be indicative of cancer and can be used in an AI based solution for early cancer diagnosis. For each of them the standard range values are: 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 work will focus on patients, both with lung cancer and non-cancer patients for control group, that have at least one of these metrics in their pathology results outside of the listed ranges.

B. Features Design

We investigated the relationship between the presence and frequency of out of range results in the pathology tests and cancer diagnosis with two groups of original features for 592 patients:

- Summary of occurrences per blood test metric
- Summary of occurrences of any metrics over a 3- or 6-month pre-diagnosis period

The features that represent summary of occurrences per blood test metric are a quantity-based feature, meaning we investigated how often the out of range result was present in patients diagnosed with cancer vs. control group patients for each of the 5 individual metrics (Platelet count, MCHC, MCV, MCH and RDW). We compared 592 lung cancer patients with 9180 non-cancer patients, with all patients having at least one metric being out of range in the period of 24 to 3 months prior diagnosis date for cancer patients and a random 21-month period for non-cancer patients. We can see from Figure 1 which depicts the total number of out of range MCV results (in percentages) for lung cancer patients vs. non-cancer group. A total of 96% of non-cancer patients showed 0 or 1 out of range value during the nearly two year period, while the cancer patients group had almost

20% with 2 or more out of range values for MCV. For both groups it is worth noticing that the number of pathology tests taken per patient vary, which is why we still have a large number of patients with 0 out of range records for the MCV, but as they have at least one of the other metrics being out of range it means we can still attempt to use the combined metrics with added feature engineering for our prediction models.

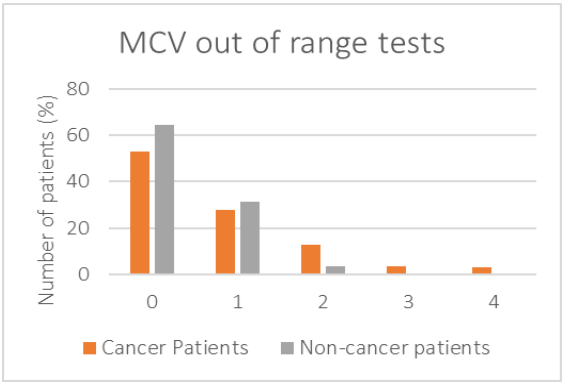

Figure 1. Number of out of range MCV tests for cancer vs. non-cancer patients.

For the features that represent summary of all the occurrences over a 3 or 6 months period, we calculated the total out of range value occurrence for all 5 of the metrics combined for the period of 24-18, 18-12,12-6 and 6-3 months before their cancer diagnosis date for cancer patients, and we selected the period of 2016-2017 for the non-cancer patients, and the same features were calculated for that period. This feature offers more of a temporal-based view of the out of range results, as it indicates how far in time we can find references between the out of range results and a cancer diagnosis. Similar to Fig. 1, Fig. 2 shows the total number of occurrences of all five metrics for the period of 6-3 months before the diagnosis date for cancer patients, which is the equivalent of a 3 month average for the non-cancer patients. We observe a similar behaviour with 95.6% of non-cancer patients showing 0 or one out of range result, while the cancer group had 23% of the patients record 2 or more out of range metrics. We discovered similar patterns in the remaining cancer metrics and time periods, and so we used these features, as well as combined versions of them as inputs for an AI model for early cancer detection. We described this in the next chapter.

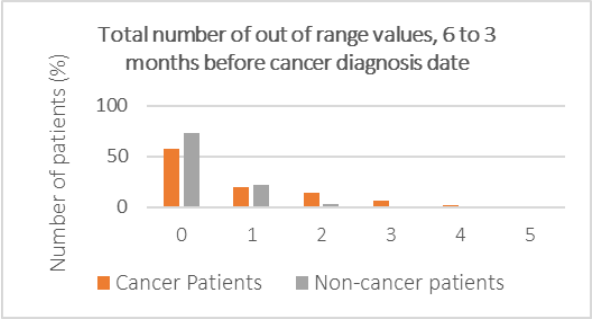

Figure 2. Number of out of range tests, 6 to 3 months before diagnosis date.

C. Deriving Additional Combinatory Features

The original 9 features described earlier (5 quantity-based and 4 temporal-based) indicate that out of range results can be related to the cancer diagnosis but they may not be sufficient to allow a Machine Learning model to distinguish between cancer and non-cancer patients. We generated additional features based on some demographics and a combination of options:

- Separating the out of range values into two separate features for upper and lower threshold out of range: some of the metrics show more out of range cases in the upper threshold of the normal range than in the lower threshold, so we want to be more specific with the type of out of range record.
- Combining quantity and temporal features: presence or absence of an individual blood test metric during a specific time frame.
- Combining the original and additional features with biological sex: several research findings indicate that male patients are known to have lung cancer at a higher rate than female [4].
- Creating age groups (groups of 10s), as majority of patients with out of range results were aged 50+. This way, we can train additional models on them too.

IV. Experiments and Results

A. Model and Samples Selection

Our research goal was to design a way to predict cancer 3 months ahead of current practice. In AI terms this translates to classification task using Machine Learning models. With the nature of the data being such with most of the features in nominal form, used models that allowed some interpretability and visualisation of the decision process: Decision Tree, AdaBoost, LightGBM and XGBoost. Decision Tree is the basic model behind all of the classifiers, with AdaBoost, LightGBM and XGBoost used to allow additional increases in performance.

The models aimed to detect potential cancer patients within a dataset that contains both cancer and non-cancer patients. This posed 2 challenges within one classification model: to correctly detect most or all of the cancer patients (true positives) and not to detect non-cancerous patients as cancer patients (false positives). We investigated True Positive Rate (TPR) False Positive Rate (FPR). We tested the performance of our models with 10 cross fold validation, with the same data used across all the models in the same fold. The number of features selected were within the range of 41-54, as our initial tests showed the performance increases for all models until around 40 features and then fluctuates very little. This provides a feature range in which we can be confident we have optimal range of results. We reported the average of the performance metrics and their respective standard deviations and used the chi-squared statistic for ranking the top features.

Another important aspect of our models was selecting the number of control group samples. With standard

patient trials, we usually have a smaller number of both patients of interest and control group patients and usually equal numbers. Our lung cancer patients dataset contains 592 patients, aged 50-100 years. We wanted to be sure that the False Positive Rate is sufficiently small, otherwise we would have a substantial number of false positives once the model is implemented in real life scenario. We investigated the ratios of control group patients vs. cancer patients of 1:1, 1.5:1 and 2:1.

B. Model Performance and Discussion

The averaged performance metrics for each of the 14 runs for the models, with 41-54 features used per run, is shown in Table I. We can observe that the performance metrics have very small standard deviations per each run with different number of features, which assures us that we can easily determine a range of features with optimal performance and allow for less performance dependence on the features selected. We used the 4 methods mentioned earlier, plus two ensembles: one standard ensemble with a voting system with OR logic (only one class 1 label is enough to assign label 1 as final), and a stack system which learns an additional model with the outputs of the other 4 models as input. The performances varied per models, with all models showing the highest True Positive Rate (TPR) for cases when the ratio of control group patients vs. cancer patients was 1:1, and the lowest False Positive Rate (FPR) for cases when the ratio was 2:1. The precision was also the highest for the 1:1 ratio, which would be a cause of concern by having too many False Positives if a 1:1 ratio was used in a real life scenario. However, the percentage of control group patients that have out of range values for the pathology metrics in relation to all potential control group patients is less than 10%, so with some additional subset generation in future work, the FPR could be very small.

The best performance is shown for the ensemble based on OR logic in the voting system, with 0.807 TPR. This means that we can provide early cancer diagnosis 3 months in advance for 8 out of 10 patients, based solely on the pathology results. Current models that focus on such task on a wide range of patients have also performed in that range, which indicates the pathology results have the potential to outperform existing models if we include additional patients' segmentation of feature engineering.

Another relevant contribution in this model is the fact that patients that have the highest mortality rate have more labels assigned in the cancer forecasted class than the non-cancer forecasted class. In our dataset, 41.9% of the patients were deceased, most within 4 years of the cancer diagnosis. With our models, the true positive forecasts contained more than 41.9% of deceased patients, meaning we detected the cancer early for the most relevant groups of patients as shown in Fig. 3. The false negative predictions on the other hand had less than 41.9% of the deceased patients, so even if we did not classify them early, there were still less high risk patients in that group. This suggests that besides providing an accurate forecast for 8 out of 10 patients in advance, we were also able to provide the forecast accurately for high risk patients whose life depend on the early forecast the most. This patient cohort is suspected to be patients with more advanced cancers and it will be subject of future research pending on additional data availability.

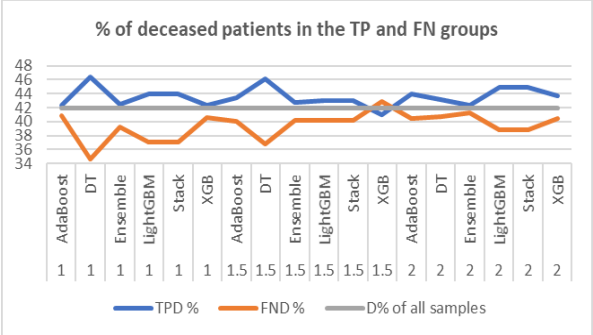

Figure 3. Percentage of deceased patients in the TP and FN groups per classification model and ratio of control group vs. cancer patients' group.

TABLE I. PERFORMANCE METRICS FOR DIFFERENT CLASSIFICATION MODELS, AVERAGED OVER 14 RUNS WITH 41-54 FEATURES

Classifier	Ratio	TPR	FPR	PPV	F1	TPR StDev	FPR StDev	PPV StDev
AdaBoost	1	**0.686**	0.278	**0.711**	**0.698**	0.025	0.007	0.007
AdaBoost	1.5	0.535	0.153	0.700	0.606	0.015	0.004	0.006
AdaBoost	2	0.394	**0.097**	0.670	0.496	0.021	0.006	0.008
DecisionTree	1	**0.613**	0.253	**0.708**	**0.657**	0.016	0.009	0.012
DecisionTree	1.5	0.549	0.218	0.626	0.585	0.013	0.007	0.006
DecisionTree	2	0.470	**0.162**	0.593	0.524	0.027	0.016	0.013
Ensemble	1	**0.807**	0.381	**0.679**	**0.738**	0.016	0.005	0.006
Ensemble	1.5	0.685	0.247	0.649	0.667	0.012	0.004	0.007
Ensemble	2	0.575	**0.181**	0.613	0.594	0.031	0.005	0.007
LightGBM	1	**0.705**	0.290	**0.708**	**0.707**	0.013	0.007	0.008
LightGBM	1.5	0.594	0.188	0.678	0.633	0.006	0.004	0.006
LightGBM	2	0.499	**0.138**	0.644	0.562	0.025	0.004	0.012
Stack	1	**0.705**	0.290	**0.708**	**0.707**	0.013	0.007	0.008
Stack	1.5	0.594	0.188	0.678	0.633	0.006	0.004	0.006
Stack	2	0.499	**0.138**	0.644	0.562	0.025	0.004	0.012
XGB	1	**0.722**	0.252	**0.742**	**0.732**	0.021	0.006	0.006
XGB	1.5	0.536	0.153	0.701	0.607	0.011	0.003	0.006
XGB	2	0.422	**0.086**	0.711	0.530	0.020	0.004	0.007

V. Conclusion

We initially set out to provide indications that some of the out of range metrics in pathology results can be indicative of cancer diagnosis. Our work shows that when assisted by AI practices, an AI based decision support tool could be implemented. This tool could be designed to be used by GPs when a patient comes to their clinic and the medical history is available to the GP. By doing so, the models can provide sufficient rationale for the GP to issue a referral for more tests or a visit to a specialist in order to confirm the diagnosis. This simple task would require no additional effort form the GP's side as the decision support would be entirely provided by the AI models, making it easy to incorporate in with current GP practices and technologies. The AI could also be used in a pathology testing clinic, providing GPs with an early cancer warning alongside the pathology results.

Not only do our models allow early indication of cancer diagnosis, they also open the opportunity to specialize per patient cohort and type of cancer – the models can be adjusted depending on patient age, location and if available cancer type and progress, allowing the patients at high risk or highest need of early accuracy access to better healthcare. Future work in this area can also provide medical insights into the specific behavior of pathology metrics in different stages of cancer, allowing for an even earlier diagnosis and treatment.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Goce Ristanoski performed the data engineering and machine learning model training. Jon Emery, Javiera Martinez Gutierrez and Damien McCarthy provided the data, medical domain knowledge and feature selection, as well as contributed to the data engineering and patient's cohort selection. Goce Ristanoski and Uwe Aickelin worked on the application of explainable AI models, statistical analysis and features design. All authors contributed to writing the paper.

References

[1] J. Corner, *et al.*, "Is late diagnosis of lung cancer inevitable? Interview study of patients' recollections of symptoms before diagnosis," *Thorax*, vol. 60, pp. 314-319, 2005.

[2] J. Hippisley-Cox and C. Coupland, "Identifying patients with suspected lung cancer in primary care: Derivation and validation of an algorithm," *British Journal of General Practice*, vol. 61, no. 592, 2011.

[3] W. Hamilton, "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*, vol. 101, pp. S80-S86. 2009.

[4] B. Iyen-Omofoman, *et al.*, "Using socio-demographic and early clinical features in general practice to identify people with lung cancer earlier," *Thorax*, vol. 68, no. 5, pp. 451-459, 2013.

[5] T. P. Hannah, *et al.*, "Mortality due to cancer treatment delay: Systematic review and meta-analysis," *BMJ*, vol. 371, 2020.

[6] M. Shapley, G. Mansell, J. L. Jordan, and K. P. Jordan, "Positive predictive values of ≥5% in primary care for cancer: Systematic review," *British Journal of General Practice*, vol. 60, no. 578, 2010.

[7] M. Bjerager, *et al.*, "Delay in diagnosis of lung cancer in general practice," *Br. J. Gen. Pract.*, vol. 56, pp. 863-868, 2006.

[8] World Health Organization. Cancer fact sheets. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cancer

[9] S. E. R. Bailey, O. C. Ukoumunne, E. Shephard, and W. Hamilton, "How useful is thrombocytosis in predicting an underlying cancer in primary care? A systematic review," *Family Practice*, vol. 34, no. 1, 2017

[10] S. E. R. Bailey, O. C. Ukoumunne, E. A. Shephard, and W. Hamilton, "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*, vol. 67, no. 659, pp. e405-e413, 2017.

[11] Victoria Cancer Council. Lung cancer. [Online]. Available: https://www.cancervic.org.au/downloads/resources/factsheets/AL1720_OCP_I-PACED_Lung_FINAL.pdf

[12] K. T. Haaf, *et al.*, "Risk prediction models for selection of lung cancer screening candidates: A retrospective validation study," *PLoS Medicine*, vol. 14, no. 4, 2017.

[13] E. L. O'Dowd, *et al.*, "What characteristics of primary care and patients are associated with early death in patients with lung cancer in the UK?" *Thorax*, vol. 70, no. 2, pp. 161-168, 2015.

[14] S. H. Bradley, M. P. T. Kennedy, and R. D. Neal, "Recognizing lung cancer in primary care," *Advances in Therapy*, vol. 36, no. 1, pp. 19-30, 2019.

[15] D. P. Weller, M. D. Peake, and J. K. Field, "Presentation of lung cancer in primary care," *NPJ Primary Care Respiratory Medicine*, vol. 29, no. 1, pp. 1-5, 2019.

[16] L. McDonald, *et al.*, "Suspected cancer symptoms and blood test results in primary care before a diagnosis of lung cancer: A case-control study," *Future Oncology*, vol. 15, no. 33, pp. 3755-3762, 2019.

[17] C. Koshiaris, *et al.*, "Early detection of multiple myeloma in primary care using blood tests: A case-control study in primary care," *Br. J. Gen. Pract.*, vol. 68, no. 674, pp. e586-e593, 2018.

[18] J. Watson, *et al.*, "Predictive value of inflammatory markers for cancer diagnosis in primary care: A prospective cohort study using electronic health records," *British Journal of Cancer*, vol. 120, no. 11, pp. 1045-1051, 2019.

Dr. Goce Ristanoski received his PhD at The University of Melbourne, Australia in 2014, specializing in Machine Learning Methods for time series forecasting. He has since worked in both industry and academic roles. His work experience covers both the research components of using Machine Learning and AI decision support tools, as well as development and production of ML based products that deliver prediction and optimization functions. He has been working in the area of applied AI for medical tasks for the past few years.

Prof. Jon Emery is the Herman Professor of Primary Care Cancer Research at the University of Melbourne, Australia, and leads a program of research on cancer in primary care with specific focus on early detection. He leads the Victorian Comprehensive Cancer Centre data linkage program which has linked clinical general practice data to hospital and clinical datasets. He has authored more than 490 papers, many in the areas of the use of AI for medical purposes in the last 20 years.

Dr. Javiera Martinez-Gutierrez is an Academic General Practitioner with post-doctoral training in cancer prevention and control. Dr. Martinez-Gutierrez is a research fellow at the University of Melbourne in the Centre for Cancer Research at the Department of General Practice. Her current work includes the development of a decision support tools to identify risk of undiagnosed cancer using the electronical medical record (EMR) in primary care. Her research focuses on early cancer detection and prevention, as well as understanding the challenges and opportunities using digital technologies in primary care research.

Dr. Damien McCarthy is a Health Data Analyst in the Department of General Practice at the University of Melbourne. He works in the Cancer in Primary Care Research Group on several projects as part of the VCCC's Data-Driven Research Program. His current work involves the large-scale linkage of primary care and hospital datasets to enable novel, longitudinal health services research across the continuum of care. This research includes analysis of linked cancer patient data at the three key points where the primary care and hospital systems intersect during a patient's cancer journey: pre-diagnosis, during active treatment and during post-treatment follow-up.

Prof. Uwe Aickelin is the Head of School of Computing and Information Systems at the University of Melbourne. Professor Aickelin has worked for more than twenty years in the fields of Artificial Intelligence, Optimisation and Data Mining, and has authored more than 280 publications. His specific expertise is in the modelling stages of problems with a focus on robust methods to overcome uncertainty.