DOI: 10.1561/9781638282372.ch13

Chapter 13

Introduction to Section III: Bringing Big Data to Oncology

By Brian Pickering

According to the World Health Organization (WHO), cancer was one of the most widespread causes of death worldwide in 2020[i] along with cardiovascular disease (ischemic heart disease and stroke). That year, it accounted for 10 million or one in six deaths. There were 2.6 million cases of breast cancer, 2.21 million of lung cancer, and 1.41 million of prostate cancer. Mortality rates are reduced with early detection and treatment, with investment for detection devoted on the one hand to appropriate and timely diagnosis and to preventative screening on the other. The rationale is simple: when cancers are identified early, they are more likely to respond favorably to treatment. This in turn has two major outcomes: first, for the patient, this increases the probability that they will survive; and second, for health services, the effective and early treatment reduces the associated care costs. Coupled with screening, especially of at-risk groups, and changes in lifestyle, the message here is clear: get in early, make an appropriate diagnosis, or take preventative action, and all will be well.

That is not the whole story, though. Notwithstanding cultural and global differences in access to healthcare, as WHO acknowledges, each cancer requires treatment via a different regimen. More importantly, though, an appropriate regimen must take into account not only the disease but also the individual patient being

i. https://www.who.int/news-room/fact-sheets/detail/cancer.

treated. Looking back at some of the most common cancers, breast cancer involving either partial (lumpectomy) or total mastectomy can have psychological consequences for gender identity. Similarly, treatments for prostate cancer which do not preserve the function of adjacent structures affect the quality of life for male patients and potentially their partners. In the following three chapters, the focus is on breast cancer (Chapter 16) and prostate cancer (Chapter 14)—one specific to women and the other to men—and the third, lung cancer (Chapter 15), affecting all genders as well as smokers and non-smokers.

Bringing together partners from industry, healthcare, and research, the breast cancer study reported here faced a very specific problem that it sought to resolve using large datasets including multimodal data routinely collected during screening. Where advanced breast cancer has developed locally, patients are often given NeoAdjuvant Chemotherapy (NAC) prior to surgery, involving chemotherapy alongside other targeted treatments. To date, clinicians have struggled to predict the outcome of NAC even where patients share similar prognostic factors. Using both clinical data (essentially the medical health record) and image data (such as MRI scans and so forth), the study reported here was able to predict four different outcomes accurately: pathologic complete response (the patient recovered) or one of three relapse states (including local reoccurrence and metastasis). The results were similar to a retrospective study using historical patient records as well as in a well-known external competitive study. At the same time, of course, this study faced all the common challenges for big data in healthcare: the sensitivity of the data as well as missing and inaccurate data. Overcoming these, however, and achieving a more encouraging prediction of outcomes for this cohort, demonstrates what can be achieved through appropriate governance and data linkage. This is big data revealing patterns in existing health records for the benefit of both service users (the women suffering a local, advanced breast cancer) and service providers (the clinicians trying to predict NAC benefit).

A naive approach to a tumor would be to remove or reduce it via chemo- or radiotherapy, surgery, or a combination thereof. Apart from the assumption that the tumor is discrete, it also fails to take into account the patient: whether they are physically or psychologically able to undergo such treatment. In the case of prostate cancer, there is a distinct danger that treatment could affect both the gender and sexual identity of the patient. Too much of the neighboring tissue is cut out, for instance, and the patient may lose sexual function or bladder control. Further, the diagnosis and surgical treatment alone calls for a MultiDisciplinary Team (MDT) to deal with oncology, urology, and radiology. It is not just the sensitivity of the medical records which is the issue here but also supporting all the required disciplines to interact and collaborate.

Led by the Karolinska Institute, one of the leading tertiary cancer care centers in Sweden, together with Philips in the Netherlands, the prostate cancer study adopted a mixed methods approach to evaluate the effects of introducing advanced technology into the clinical setting (i.e., prediction based on big data) as well as patient and clinician responses to that technology. Across both retrospective and prospective cohorts, the team looked at the quality of the MDT conferences and patient satisfaction with the introduction of the predictions into the decision-support environment. Predictions were generated by modeling the big datasets and integrating them into Philips's visualization environment. Positive outcomes were reported, especially for the clinicians. Of course, they too faced the common issue of data quality—including setting themselves a particular challenge in attempting to integrate an analysis of freeform text generated across many different hospitals—and the de facto sensitivity of the data. What they have demonstrated though is the potential for predictive modeling in the delivery of effective cancer treatment which by its nature requires collaboration between colleagues from different specializations to deliver the maximum benefit to the patient.

Despite obvious advances, lung cancer treatments are still not personalized and can be associated with adverse effects in consequence. One particular challenge is identifying how long to engage with chemotherapy. By its very nature, chemotherapy is toxic, and identifying outcomes, especially for atypical patient groups, is difficult. Yet, as with many areas of medicine and clinical care, there is not only a significant amount of historical as well as current patient data such as electronic health records but also a substantial research literature. If treatment planning is based only on the specific health records of the patient, though, it would be difficult to identify broader trends that may help patient care going forward. More importantly, perhaps, it would also mean that what appears an exceptional or rare case in one hospital is not seen against the context of potentially many others reported in the literature. Given enough time and resources, clinicians may well be able to consult the literature or other experts. Chapter 19 (Implementation and Impact of AI for the Interpretation of Lung Diseases in Chest CTs), for instance, explores one way of searching for relevant image data to support radiologists. In the lung cancer study, the scope is broadened out to help the clinician visualize what is in the literature and to exploit such knowledge in the service of not only diagnosis but also preventative screening.

The study demonstrates the integration of multiple and disparate data sources into a single knowledge base (in this case, a knowledge graph) representing an ontology of one and a half million triplets associated with lung cancer incidence, potential indicators via emergency room visits, and adverse, oncologically related drug interactions. Data sources include not only traditionally sensitive healthcare records but also open-source reports from the academic literature. Exploiting the latter, of course, required Natural Language Processing (NLP) capabilities. In the prostate cancer study, the effectiveness of NLP was compromised by the variability in freeform text from different teams in different institutions. Here though, the NLP has been shown to provide real benefit in the more constrained, though more verbose text of academic literature. As a result, clinicians can then query the knowledge graph—even via freeform text queries—in support of their own treatment planning as well as screening activities.

The three studies in this section deal with some of the most frequent cancers affecting women, and men as well as both. Despite the obvious challenges of data sensitivity and the consequent governance structures that must be respected (Chapter 4—Lessons Learned in the Application of the General Data Protection Regulation to the BigMedilytics Project) and the common problems of data quality (Chapter 25—Data Processing in Healthcare Using CRISP), the three studies have demonstrated the real practical potential for big data in the fight against cancer. Integrating data from different sources, in different formats, and providing different types of information relevant to the specific case, these studies show what can be done with big data in the provision of healthcare services at various stages of the healthcare lifecycle. Perhaps, more importantly, in supporting the patient and optimizing their outcomes, but also for the clinicians, the use of big data here facilitates cross-disciplinary collaboration and consulting multiple sources. Like the other chapters in this volume, these oncology studies have moved beyond the assumed problems with big data exploitation in healthcare to demonstrate real potential across the sector.