frontiers | Frontiers in Bioinformatics

TYPE Opinion
PUBLISHED 08 February 2023
DOI 10.3389/fbinf.2023.1112113

Visualising disease trajectories from population-wide data

Jessica Xin Hjaltelin[1], Hannah Currant[1], Isabella Friis Jørgensen[1] and Søren Brunak[1,2]*

[1]Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark, [2]Copenhagen University Hospital, Rigshospitalet, Copenhagen, Denmark

Check for updates

OPEN ACCESS

EDITED BY
Sean O'Donoghue,
Garvan Institute of Medical Research,
Australia

REVIEWED BY
Burcu Bakir-Gungor,
Abdullah Gül University, Türkiye

*CORRESPONDENCE
Søren Brunak,
✉ soren.brunak@cpr.ku.dk

SPECIALTY SECTION
This article was submitted to
Data Visualization,
a section of the journal
Frontiers in Bioinformatics

RECEIVED 30 November 2022
ACCEPTED 17 January 2023
PUBLISHED 08 February 2023

CITATION
Hjaltelin JX, Currant H, Jørgensen IF and Brunak S (2023), Visualising disease trajectories from population-wide data. *Front. Bioinform.* 3:1112113.
doi: 10.3389/fbinf.2023.1112113

KEYWORDS

disease trajectory, visualisation, networks, personalised medicine, pancreatic cancer, comorbidity, multimorbidity, sankey

Introduction

The rise of precision medicine as both a research discipline and clinical practice looks to improve clinical experience by using the wealth of health-associated data to tailor treatments towards the individual patient. Such data includes electronic health records alongside omics data that are increasingly collected in healthcare settings including genetic data, proteomics, metabolomics and more. Digitalisation of electronic healthcare has led to a wealth of data describing individuals' health and disease status. Taken together, such data that describes individuals molecular biology and pathology is being used to expand our understanding and utility of personalised medicine.

Using electronic healthcare records and disease histories for either explorative or predictive purposes can help identify risk factors and stratify patients by disease risk which can ultimately inform screening protocols. In the recent decade, studies have been utilising the concept of disease trajectories in classical statistical approaches to explore risk factors and complications, deep learning algorithms for disease onset prediction or patient stratification, amongst others (Jensen et al., 2017; Shickel et al., 2018; Hu et al., 2019; Lademann et al., 2019; Nielsen et al., 2019; Thorsen-Meyer et al., 2020; Placido et al., 2022). Many studies have previously analysed diseases in an either mono- or bidirectional manner. Today, trajectory analyses and visualisations can utilise temporal information in expanding large health data sets allowing for consideration of comorbidities for different patients. Comorbidity and multimorbidity refer to the presence of more than one disease in a single patient and has been increasingly recognised as a crucial consideration when diagnosing and treating patients (Hu et al., 2016).

Visualisation can be a powerful tool for understanding all steps in an analysis using large data sets across a temporal axis. Denmark is one of the leading countries in collecting decades of longitudinal population-wide health data. Denmark has a wealth of health registries for which patients can be linked on an individual level through the unique Central Person Register (CPR) identifier. One of the largest and most comprehensive national registries is the Danish National Patient Registry (NPR), which covers around 8.2 million Danish patients over nearly 45 years. The visualisation of this type of large health dataset can be a highly complex matter. Here, we will use pancreatic cancer as an example to visualise temporal disease patterns in the NPR, giving examples of different types of plots and tools useful for the overviewing and analysing large longitudinal health data.

Pancreatic cancer is one of the most lethal cancer types with a 5-year survival rate at only 8% (American Cancer Society, 2020) and has been estimated to become the second leading cause of cancer in 203010. Due to the lack of clear symptoms, this cancer type is often diagnosed at a later stage resulting in poor outcomes. Hence, the need for detecting early symptoms and risk factors is crucial. We will highlight some of the key forms of data visualisation methods utilised

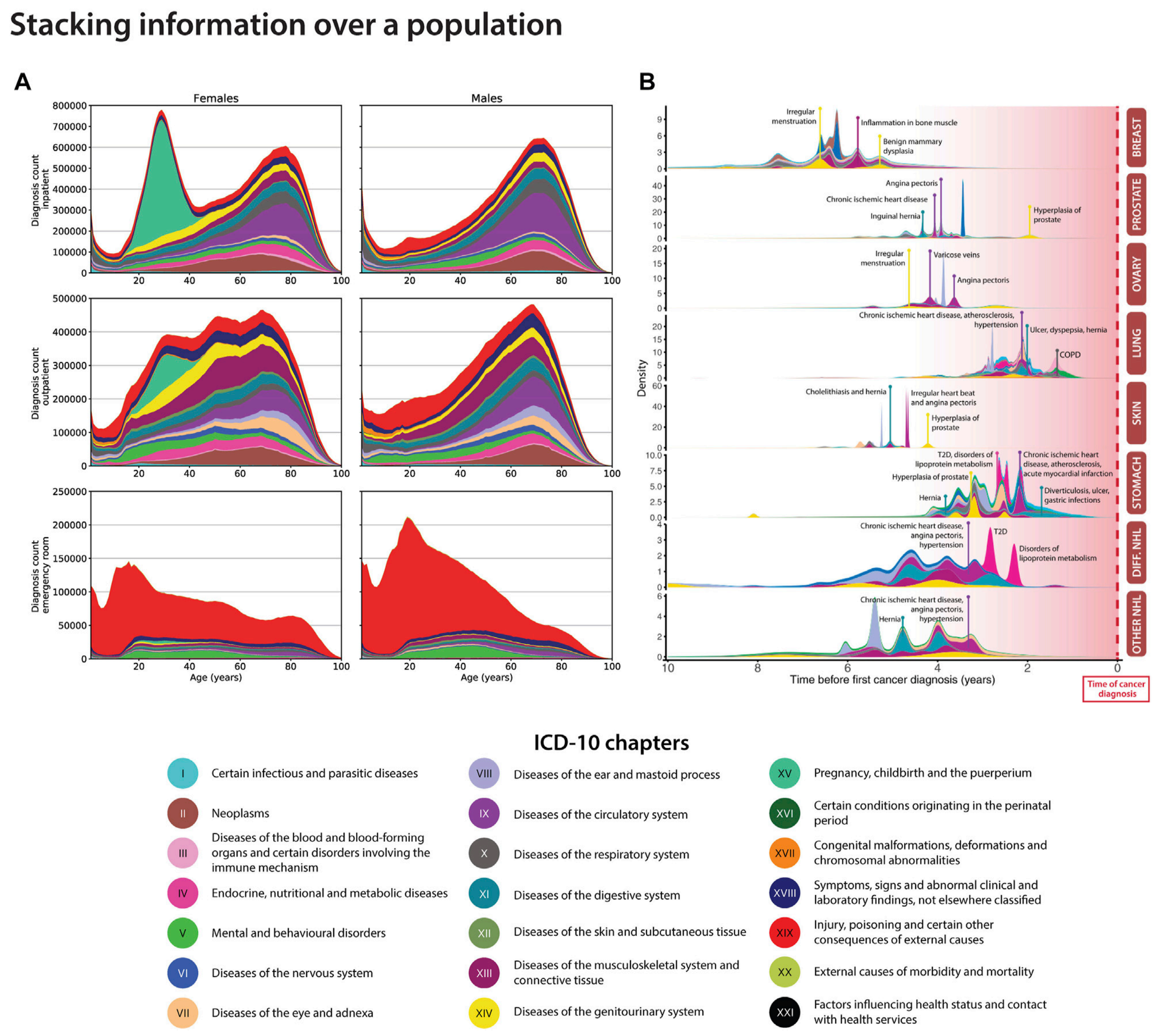

FIGURE 1
Density plots for visualising health data for the Danish population. **(A)**. The stacked density plot compares the amount of ICD-10 chapters between sexes for the entire Danish population *via* the National Patient Registry (NPR) (Adapted from Jensen et al. (2014). **(B).** The stacked density plot shows the occurrences of significant correlated disease pairs previous to top 8 cancer types (Hu et al., 2019). Breast: breast cancer; Prostate: prostate cancer; Ovary: ovarian cancer; Lung: lung cancer; Skin: skin cancer, Stomach: stomach cancer; Diff. NHL: diffuse non-Hodgkin's lymphoma; Other non-Hodgkin's lymphoma.

in the analysis of disease trajectories, the ways in which these enrich our understanding of the data and the conclusions drawn.

Visualising a population

Visualisation is a powerful tool in understanding the cohort or population in which all analysis will take place, and to compare it to other populations. This visualisation can have the two-fold value: Checking that the data behaves as expected and is of high-quality; Identify novel patterns that may be indicative of unique biology or disease mechanism. For example, in the analysis of electronic healthcare data, an initial analysis may be to look at the distribution of different diagnosis types across age (Figure 1A). In Denmark, disease diagnoses are registered nation-wide in the NPR by the International Classification of Diseases version 8 and most recently (Rahib et al., 2014) (ICD-10). These are coded in electronic registries alongside the date of the diagnosis and patient birth information, from which the age of diagnosis can be derived. This can be plotted as a stacked density plot, and further stratified by sex and age (Figure 1A, ICD-10 period only) (Jensen et al., 2014). From Figure 1A we are able to notice general overview trends for the cohort such as the pregnancy chapter, emergency room contacts at younger ages and the increasing cardiovascular diagnoses from age 60. Stacked density plots can also be used to gain an overview of comorbidities along a temporal axis that represents a relative time since diagnosis of interest. For example, Figure 1B shows an overview of significant diagnoses (coloured by ICD-10 chapters) in the years up till a cancer diagnoses. This allows for

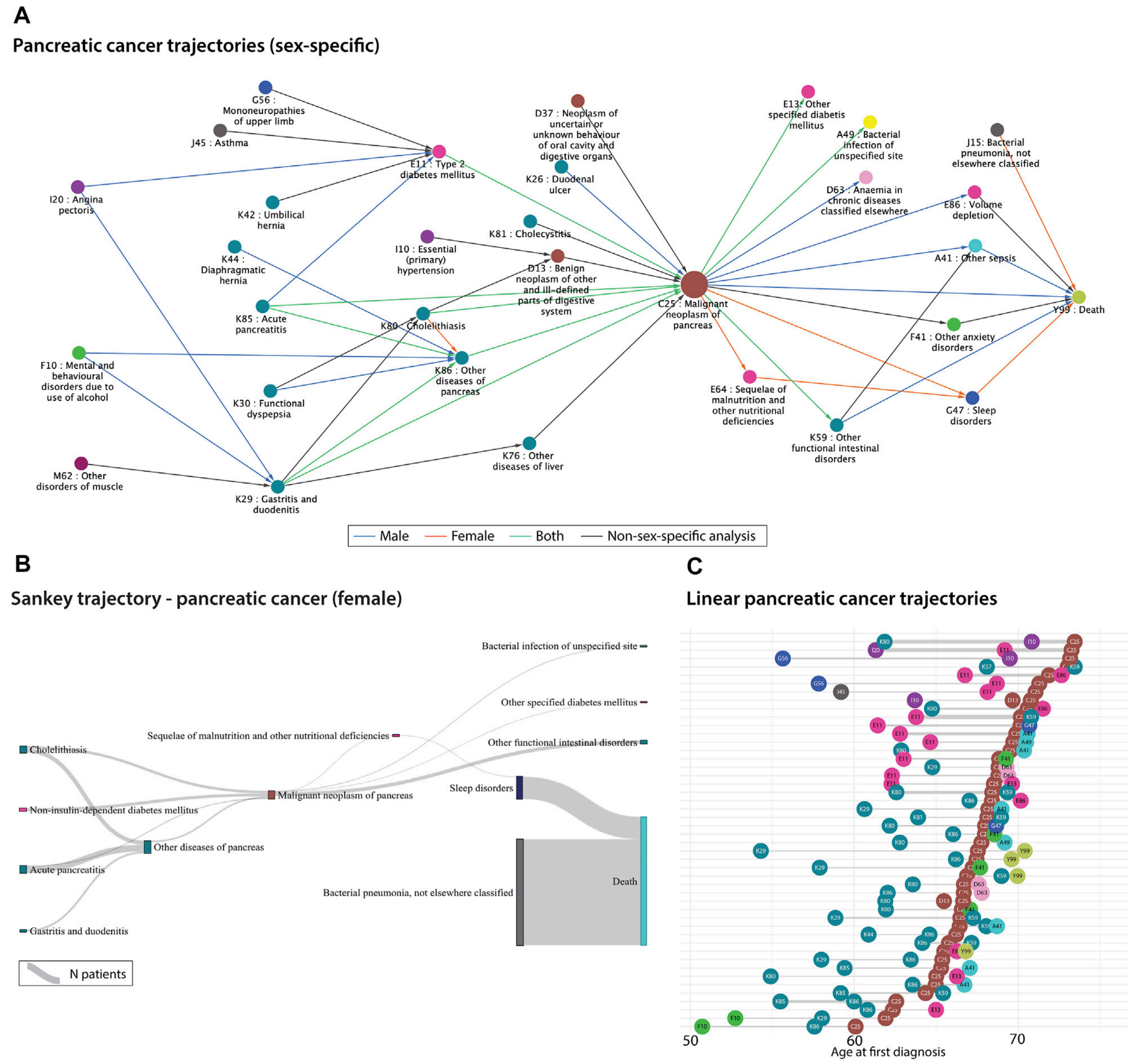

FIGURE 2
(A). Sex-specific significant pancreatic cancer trajectories extracted from 7.2 million Danish disease histories from the National Danish Patient Registry (NPR). The correlations were extracted from the Danish Disease Trajectory Browser (DTB) (Siggaard et al., 2020). Edges are coloured according to sex-specific analyses run by the DTB. **(B)**. Sankey plot of the significant female-specific pancreatic cancer trajectories generated based on 7.2 million Danish disease histories from the National Danish Patient Registry (NPR). The disease-disease correlations were extracted from the publicly available Danish Disease Trajectory browser (Siggaard et al., 2020). The thickness of the links represents the number of patients. **(C)**. Linear trajectories plot showing 47 of the significant pancreatic cancer trajectories with three consecutive diseases. The temporal order of diagnoses is here represented as the mean age at the first diagnosis for patients that follow each of the trajectories. The thickness of the links represents the number of patients. The colours from all three figures represent the ICD-10 chapters (see Figure 1 legend).

the mapping of potential risk factors on a time scale, gaining a temporal trajectory. For example, we are able to see that as one might expect, in both breast cancer and ovarian cancer, irregular menstruation is observed in numerous patients prior to the diagnosis. Further, in cancers including that of the stomach and diffuse large B cell lymphoma (Diff. NHL), we observe a type 2 diabetes (T2D) diagnosis prior to the cancer diagnosis. This summarises not only disease pairs across a nation-wide cancer landscape, but also visualises them on a temporal scale prior to the event of interest.

The disease trajectory highway and temporality

Disease trajectories are longitudinal sequences of diseases that occur in a temporal order. Diseases could for example be represented by ICD-10 codes, symptom codes, text mined disease codes or symptoms (Jensen et al., 2017), (Jensen et al., 2014; Beck et al., 2016; Siggaard et al., 2020). The temporality of diseases can be very useful to stratify patients into different risk groups, understand comorbidities and multimorbidities or improve disease progression patterns. Examples of how to visualise diseases using a network view could be *via* the Cytoscape software (Shannon et al., 2003) or the Danish Disease Trajectory Browser (for Danish disease correlations) (Siggaard et al., 2020). For the latter, population-wide summarised data from the NPR can be collected to visualise significant disease trajectories for a disease of interest. Figure 2A shows a network extracted from the Danish Disease Trajectory Browser for pancreatic cancer patients. The network nodes are coloured in chapters according to the ICD-10 chapters and edges are coloured according to sex-specific disease-disease correlations. For this example, we can see that male-specific correlations involve angina pectoris and alcohol abuse disorders, while female patients have post-cancer malnutrition deficiencies and sleep disorders. Although the directional correlations are significant in the analysis (Siggaard et al., 2020), they have not been proven to be causal.

Another useful visualisation method for patient or disease trajectories are Sankey and alluvial flow diagrams. The width of the diagram bars conveys the number of patients in a specific link. We used publicly available pancreatic cancer-specific disease correlations from the Danish Disease Trajectory Browser (for Danish disease correlations) (Siggaard et al., 2020) to visualise patient groups flowing across disease states. One should be aware that alluvial and Sankey diagrams have different underlying assumptions. One difference is for example that alluvial diagrams have aligned bars in columns/dimensions, whereas the bars in Sankey plots can be distributed anywhere, depending on the specific Sankey algorithm. In Figure 2B, trajectories are combined by disease pairs using significant pancreatic cancer disease pairs from the DTB. Here, the disease pairs have been linked by the Sankey algorithm, thus it may not be the same patient group that traverses an entire trajectory. These types of flow diagrams are getting more focus for visualising longitudinal healthcare data such as prescription trajectories (Aguayo-Orozco et al., 2021), symptom trajectories (Lademann et al., 2019), cancer trajectories (Hu et al., 2019), hospital flow from acute coronary syndrome (Pinaire et al., 2021) etc.

Disease trajectory networks can be useful to get an overview of the alternative disease routes for patients. Even though, the trajectories are temporal, Figures 2A,B do not inform about the time between the diagnoses. Figure 2C visualises single linear disease trajectories as a function of time. It is represented as the mean age at the first diagnosis for the patients following the specific trajectory and thereafter, sorted using the mean age of pancreatic cancer. Thus, one can investigate the time between diagnoses and the average age of diagnosis. The pancreatic cancer diagnosis often appears after the age of 60 which is consistent with a late diagnosis. For example, diseases from "diseases of the digestive system" chapter (cyan coloured nodes) seem to appear earlier than type II diabetes (pink node E11) in relation to pancreatic cancer, which could be valuable to consider when developing screening protocols or tools. The poor prognosis of pancreatic cancer is also visualised here, since some trajectories include death shortly after the pancreatic cancer diagnosis. Disease trajectories can be combined with mortality information to stratify patients and optimally improve treatment or surveillance for these patient groups (Beck et al., 2016; Shang et al., 2022; Yang et al., 2019).

Final considerations

With the constantly increasing amounts of data within healthcare and research, there is a huge need for improved and more dynamic and interactive visualisation tools. Most visuals today are static images. But the complexity of data that expands by both velocity, variety and volume, needs new methods for comprehending, analysing and interpreting them in the multidimensional spaces they live in.

Increasingly, studies are using disease trajectories together with deep learning models for risk prediction and stratification of patients. Here, a big challenge is to visualise and explain temporal patterns picked up by these models, which is essential for applying them to decision-making in the clinics. Currently, some tools have been developed to target this problem including SHAPley values, deepExplain and others (Lundberg and Lee, 2017). Although the "static" networks shown above visualises data from a certain time interval only, another task will be to develop models and visualisation of patient's disease progression in real time. This is particularly important within intensive care, where the data richness is much higher than in the diagnosis trajectories shown here. Due to the emergence of wearable data all patients will with time grow in data richness begging the development of live models of high complexity.

Author contributions

JXH has drafted the manuscript and created all the figures, except Figure 2C. HC and IFJ helped draft the manuscript and IFJ created Figure 2C. SB critically revised the manuscript.

Funding

We would like to acknowledge funding from the Novo Nordisk Foundation (grant agreements NNF14CC0001 and NNF17OC0027594), the Braindrugs (R279-2018-1145) as well as ImmunAID (779295). HC was funded by the European Union's Horizon 2020 research and innovation program under Marie Sklodowska-Curie Actions grant no. 101064250, RMCmplxPheno. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or Horizon Europe. Neither the European Union nor the granting authority can be held responsible for them. We acknowledge Roc Reguant for contributing to visual details in Figure 1A.

Conflict of interest

SB has ownerships in Intomics A/S, Hoba Therapeutics Aps, Novo Nordisk A/S, Lundbeck A/S, ALK-Abello A/S and managing board memberships in Proscion A/S and Intomics A/S outside the submitted work. All other authors declare no competing interests.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Aguayo-Orozco, A., Haue, A. D., Jorgensen, I. F., Westergaard, D., Moseley, P. L., Mortensen, L. H., et al. (2021). Optimizing drug selection from a prescription trajectory of one patient. *NPJ Digit. Med.* 4, 150. doi:10.1038/s41746-021-00522-4

American Cancer Society (2020). *Cancer facts & figures 2020*. Atlanta, Ga: American Cancer Society.

Beck, M. K., Jensen, A. B., Nielsen, A. B., Perner, A., Moseley, P. L., and Brunak, S. (2016). Diagnosis trajectories of prior multi-morbidity predict sepsis mortality. *Sci. Rep.* 6, 36624. doi:10.1038/srep36624

Hu, J. X., Helleberg, M., Jensen, A. B., Brunak, S., and Lundgren, J. (2019). A large-cohort, longitudinal study determines precancer disease routes across different cancer types. *Cancer Res.* 79, 864–872. doi:10.1158/0008-5472.can-18-1677

Hu, J. X., Thomas, C. E., and Brunak, S. (2016). Network biology concepts in complex disease comorbidities. *Nat. Rev. Genet.* 17, 615–629. doi:10.1038/nrg.2016.87

Jensen, A. B., Moseley, P. L., Oprea, T. I., Ellesoe, S. G., Eriksson, R., Schmock, H., et al. (2014). Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients. *Nat. Commun.* 5, 4022. doi:10.1038/ncomms5022

Jensen, K., Soguero-Ruiz, C., Oyvind Mikalsen, K., Lindsetmo, R. O., Kouskoumvekaki, I., Girolami, M., et al. (2017). Analysis of free text in electronic health records for identification of cancer patient trajectories. *Sci. Rep.* 7, 46226. doi:10.1038/srep46226

Lademann, M., Lademann, M., Boeck Jensen, A., and Brunak, S. (2019). Incorporating symptom data in longitudinal disease trajectories for more detailed patient stratification. *Int. J. Med. Inf.* 129, 107–113. doi:10.1016/j.ijmedinf.2019.06.003

Lundberg, S., and Lee, S.-I. (2017). *A unified approach to interpreting model predictions*. arXiv [cs.AI].

Nielsen, A. B., Thorsen-Meyer, H. C., Belling, K., Thomas, C. E., and Chmura, P. J. (2019). Survival prediction in intensive-care units based on aggregation of long-term disease history and acute physiology: A retrospective study of the Danish national patient registry and electronic patient records. *Lancet Digital Health* 1, e78–e89. doi:10.1016/s2589-7500(19)30024-x

Pinaire, J., Aze, J., Bringay, S., Poncelet, P., Genolini, C., and Landais, P. (2021). Hospital healthcare flows: A longitudinal clustering approach of acute coronary syndrome in women over 45 years. *Health Inf. J.* 27, 146045822110330. doi:10.1177/14604582211033020

Placido, D., Yuan, B., Hjaltelin, J. X., Haue, A. D., Chmura, P. J., Yuan, C., et al. (2022). *Pancreatic cancer risk predicted from disease trajectories using deep learning*. 10.1101/2021.06.27.449937.

Rahib, L., Smith, B. D., Aizenberg, R., Rosenzweig, A. B., Fleshman, J. M., and Matrisian, L. M. (2014). Projecting cancer incidence and deaths to 2030: The unexpected burden of thyroid, liver, and pancreas cancers in the United States. *Cancer Res.* 74, 2913–2921. doi:10.1158/0008-5472.can-14-0155

Shang, X., Zhang, X., Huang, Y., Zhu, Z., Zhang, X., Liu, S., et al. (2022). Temporal trajectories of important diseases in the life course and premature mortality in the UK Biobank. *BMC Med.* 20, 185. doi:10.1186/s12916-022-02384-3

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res.* 13, 2498–2504. doi:10.1101/gr.1239303

Shickel, B., Tighe, P. J., Bihorac, A., Rashidi, P., and Deep, E. H. R. (2018). Deep ehr: A survey of recent advances in deep learning techniques for electronic health record (ehr) analysis. *IEEE J. Biomed. Health Inf.* 22, 1589–1604. doi:10.1109/jbhi.2017.2767063

Siggaard, T., Reguant, R., Jorgensen, I. F., Haue, A. D., Lademann, M., Aguayo-Orozco, A., et al. (2020). Disease trajectory browser for exploring temporal, population-wide disease progression patterns in 7.2 million Danish patients. *Nat. Commun.* 11, 4952. doi:10.1038/s41467-020-18682-4

Thorsen-Meyer, H.-C., Nielsen, A. B., Nielsen, A. P., Kaas-Hansen, B. S., Toft, P., Schierbeck, J., et al. (2020). Dynamic and explainable machine learning prediction of mortality in patients in the intensive care unit: A retrospective study of high-frequency data in electronic patient records. *Lancet Digital Health* 2, e179–e191. doi:10.1016/s2589-7500(20)30018-2

Yang, H., Pawitan, Y., He, W., Eriksson, L., Holowko, N., Hall, P., et al. (2019). Disease trajectories and mortality among women diagnosed with breast cancer. *Breast Cancer Res.* 21, 95. doi:10.1186/s13058-019-1181-5