Cancer subtypes in aetiological research

Cancer subtypes in aetiological research

Lorenzo Richiardi[1,2] · Francesco Barone-Adesi[3] · Neil Pearce[4,5]

& Lorenzo Richiardi
lorenzo.richiardi@unito.it

[1] Cancer Epidemiology Unit, Department of Medical Sciences, University of Turin and CPO-Piemonte, Turin, Italy

[2] Harvard T.H. Chan School of Public Health, Boston, MA, USA

[3] Department of Pharmaceutical Sciences, University of Eastern Piedmont, Novara, Italy

[4] Department of Medical Statistics, London School of Hygiene and Tropical Medicine, London, UK

[5] Centre for Public Health Research, Massey University, Wellington, New Zealand

Abstract Researchers often attempt to categorize tumors into more homogeneous subtypes to better predict prog-nosis or ~~understand pathogenic mechanisms. In clinical research, typic~~ally the focus is on prognosis: the tumor subtypes are intended to be associated with specific responses to treatment and/or different clinical outcomes. In aetiological research, the focus is on identifying distinct pathogenic mechanisms, which may involve different risk factors. We used directed acyclic graphs to present a framework for considering potential biases arising in aeti-ological research of tumor subtypes, when there is incomplete correspondence between the identified subtypes and the underlying pathogenic mechanisms. We identified two main scenarios: (1) weak effect, when the tumor sub-types are identified through combinations of characteristics and some of these characteristics are affected by factors that are unrelated with the underlying pathogenic mecha-nisms; and (2) lack of causality, when the set of charac-teristics corresponds with a mechanism that is actually not a cause of the tumor of interest. Examples of the magnitude of bias that can be introduced in these situations are provided. Although categorization of tumors into homogenous subtypes may have important implications for aetiological research and identification of risk factors, the characteristics used to classify tumors into subtypes should be as close as possible to the actual pathogenic mechanisms to avoid interpretative biases. Whenever our knowledge of these mechanisms is limited, research into risk factors for tumor subtypes should first aim to causally link the char-acteristics to the pathogenic mechanisms.

Keywords Cancer subtypes Molecular characteristics Bias Disease classification Aetiological research

Introduction

Clinical and aetiological research often aims to categorize disease entities into more homogeneous subtypes to better predict prognosis, or to improve the understanding of pathogenic mechanisms. Typically, a disease is classified into subtypes that share specific characteristics. In partic-ular, any characteristic associated with prognosis and/or response to treatment may identify subtypes that differ regarding their clinical outcome; however, the same char-acteristics may not correspond to specific aetiologic mechanisms, and vice versa.

Research into the identification of disease subtypes has recently increased due to advances in high-throughput techniques to explore molecular patterns, the availability of large biological databases, and the implementation of col-laborative initiatives to obtain and analyze large numbers of tissue samples in a standardized fashion. This increased research activity is tightly linked with the concept of pre-cision medicine which involves the identification of disease subtypes to better target the treatment, estimate prognosis

and/or manage clinical follow-up. This concept is having a particular impact on clinical oncology and cancer research [1], but is being adopted also in non-malignant diseases, including, for example, asthma [2], cardiovascular disease [3] and diabetes [4].

A related development is to use disease subtypes in aetiological studies. For example, researchers have identi-fied subtypes of breast cancer, based on the expression of ESR1 (estrogen receptor), PGR (progesterone receptor), and ERBB2 (HER2). Breast cancers with higher expression of ESR1 or PGR can be sub-classified into luminal A or luminal B depending on whether the expression of prolif-erative genes (ERBB2 or ki67) is low or high; non-luminal breast cancers can be classified as ERBB2? or triple negative, which in turn may be further classified on the basis of expression of basal cytokeratins or epidermal growth factor receptor [5]. Additional markers are being evaluated using more in-depth molecular profiling [6]. While ESR1, PGR and ERBB2 are usually tested for prognosis and to inform therapeutic choices [7], it has been recently proposed to consider breast cancer subtypes for etiological purposes. Epidemiological studies have shown that triple negative breast cancer may have specific risk factors [8].

Traditionally, tumor classification has been primarily based on the anatomical site of origin of the tumor [9]. Then, within each tumor site, stratification is typically carried out on the basis of the cell type of origin and, sometimes, other specific characteristics. It is now becoming increasingly possible to further subdivide the tumors on the basis of their molecular features [10]. Thus the process of identification of tumor subtypes currently goes from the identification of an organ-specific tumor (e.g. ‘‘breast cancer’’), to the classification of that tumor into specific subtypes (e.g. ‘‘triple-negative breast cancer’’ or even finer subtypes [6, 11]).

It is debatable whether such tumor subtypes represent subtypes of a single disease, or should be treated as dif-ferent diseases. We will approach this issue pragmatically assuming that an organ-specific tumor type has been identified, and the issue is whether and how to divide this into subtypes. This approach mimics the current clinical approach to tumor heterogeneity: the characterization of tumor subtypes—often based on molecular analyses of the tumor tissue—logically follows the diagnosis of the tumor as an organ-specific entity. For example, a patient is first diagnosed with a breast cancer and, then, after having the results of immunohistochemistry assays on the tumor tis-sue, that tumor is further subclassified on the basis of the receptor status.

We support the need to reach a clearer definition of tumors in terms of pathogenic mechanisms. However, in this paper we argue that aetiological research into tumor

subtypes may, under some circumstances, be problematic. A better understanding of these limitations is required for appropriate planning and interpretation of aetiological studies based on tumor subtypes. It is important to note that prognostic research and aetiological research into cancer subtypes have fundamental differences. In prognostic research any characteristic that is strongly associated with the clinical outcome of interest (e.g. response to treatment or mortality) may be effectively used to classify the patients into risk categories and the cancers into subtypes. There is no need to identify distinct causal pathways related to the distinct subtypes and the clinical outcome can be observed and assessed in classical epidemiological studies. Conversely, in aetiologic research the focus is on causal pathways and the subtypes should identify distinct (unobserved) pathogenic mechanisms. In this paper, we first present a conceptual framework for considering the correspondence between pathogenic mechanisms and the identification of the disease subtypes. We then assess two scenarios in which there is not a direct correspondence between characteristics and distinct pathogenic mecha-nisms: (1) weak effect, when the tumor subtypes are identified through combinations of characteristics and some of these characteristics are affected by factors that are unrelated with the underlying pathogenic mechanisms; (2) lack of causality, when the set of characteristics corre-sponds with a mechanism that is actually not a cause of the tumor of interest. Finally we provide some suggestions on how to conduct research to establish the links between the characteristics and the pathogenic mechanisms.

Tumor subtypes and pathogenic mechanisms

In the context of aetiological research, tumor subtypes have aetiological value if they identify distinct pathogenic mechanisms. In Fig. 1 we use a directed acyclic graph (DAG) to depict the supposed causal relationships between the diagnosis of a tumor, the identification of the tumor subtypes, and the associated pathogenic mecha-nisms: E denotes an exposure, A denotes a pathogenic mechanism causing an organ-specific tumor T and C is a characteristic that is used alone or in combination with other characteristics to define the tumor subtypes S once the tumor T has been diagnosed. Both pathogenic mech-anisms (A_1, A_2) lead to the development of the tumor T. However tumors caused by the two mechanisms will show different characteristics (C_1 and C_2), which are then used to define the disease subtypes S. It is also possible that the characteristics used to define the tumor subtypes contribute to the initial diagnosis of the tumor, in which case there would be arrows from C_1 and C_2 to T. How-ever, in the current example, we assume that the

characteristics that are used to define the subtypes are not used to reach the first diagnosis, but are evaluated after the diagnosis of the tumor.

For simplicity, in Fig. 1 and throughout the article we assume that A and C variables are binary. It is thus assumed that a tumor can have two subtypes: the first is due to A_1 (and identified through C_1), the second is due to A_2 (and identified through C_2). If it is biologically possible that A_1 and A_2 coexist (as, for example, in the case of intratumor heterogeneity), then the two subtypes can also coexist. A simple example of the scenario described in Fig. 1 is the sub-classification of tumors on the basis of the cell type of origin. For example germ-cell testicular cancer can be divided into seminomas and nonseminomas: these are biological and, potentially, clinical different subtypes, and it is not uncommon that a seminoma and a nonsemi-noma coexist in the same patient.

In the model proposed in Fig. 1, the subtypes directly correspond to pathogenic mechanisms. The exposures may act specifically on a pathogenic mechanism (E_1 or E_3) or may be shared by the different pathogenic mechanisms (E_2). It should be noted that C_1 and C_2 may be a set of characteristics (rather than single characteristics), which may even influence each other, but should still be specific for a single pathogenic mechanism. Furthermore, the dia-gram depicted in Fig. 1, as well as the other diagrams that will be used in the manuscript, are simplified models that cannot fully depict the complexity of the molecular mechanisms involved in the pathogenic mechanisms lead-ing to a disease. We propose these diagrams only as frameworks for discussing possible biases arising in the etiological research of cancer subtypes. To illustrate the implications of having such a framework, we can take the example of germ-cell testicular cancer. Studies of the aetiologcal heterogeneity of seminomas and nonseminomas typically classify mixed tumors (i.e. those with both seminoma and non-seminoma components) together with nonseminomas. This classification is based only on prog-nostic considerations, as mixed tumors have a similar prognosis and response to treatment as that of pure non-seminomas. However, on the basis of the diagram reported in Fig. 1, we would have to classify mixed tumors together with seminomas when we analyze risk factors for semi-nomas and together with nonseminomas when we analyze risk factors for this cancer subtype. Alternatively, we could exclude mixed tumors from the study, but aetiologically, there is no reason to group mixed tumors only with nonsemimonas.

Problems of causal interpretation

It is possible that the assumption of direct correspondence between the characteristics used to define the subtypes and the pathogenic mechanism is not valid. This may affect the interpretability of aetiological studies of tumor subtypes. We discuss two scenarios, which are shown, respectively, in Fig. 2 (hereafter labeled as ‘‘weak effect’’) and Fig. 3 (hereafter labeled as ‘‘lack of causality’’).

Scenario 1: weak effect

In the scenario depicted in Fig. 2a, the set of characteristics C_2 is caused by both pathogenic mechanisms A_1 and A_2, thus a subtype is defined by the presence of C_1 and C_2, while the other subtype has only C_2 features. This scenario may apply whenever the tumor subtypes are defined by combinations of characteristics (the definition of breast cancer subtypes, for example, requires the combination of ESR1 and PGR expression with ERBB2 expression). The situation depicted in Fig. 2a would lead to correct causal interpretations. However there may be problems of causal interpretation if, as shown in Fig. 2b, some of the charac-teristics used to define the tumor subtypes are affected by factors that are unrelated with the pathogenic mechanisms. In Fig. 2b, the exposure E_4 affects the characteristic C_1 without acting on the pathogenic mechanism A_1. Since C_1 is explained also by E_4, the pathogenic mechanism A_1 is not a necessary and sufficient cause of C_1 and thus its effect on C_1 is weakened (hence the label weak effect).

The scenario described in Fig. 2b, in which the char-acteristic C_1 is caused by both the pathogenic mechanism A_1 and the exposure E_4, may imply that: (1) some individuals have the set of characteristics C_1 for mecha-nisms that are unrelated with the pathogenic mechanism A_1; or that (2) the pathogenic mechanism A_1 does not always produce the associated characteristics C_1. These two mechanisms are not mutually exclusive. Their conse-quence is that individuals who experience the pathogenic mechanism A_2 can have both C_1 and C_2 and be spuriously attributed to the pathogenic mechanism A_1, and, vice versa, individuals who experienced the pathogenic mechanism A_1 can have only C_2 and be spuriously attributed to the pathogenic mechanism A_2. If all variables are binary (0 = absent, 1 = present): (1) $C_1 = 1$ does not imply necessarily that $A_1 = 1$, as presence of C_1 could be due also to $E_4 = 1$, and (2) $C_1 = 0$ does not imply necessarily that $A_1 = 0$, for example if there is an interaction between E_4 and A_1 that produces $C_1 = 0$ when, $A_1 = 1$ and $E_4 = 1$. It follows that presence of C_1 may occur in absence of A_1 and absence of C_1 may occur in presence of A_1. In a given study, if E_4 were known and observed one could consider how to treat this variable to decrease the impact of the weak effect problem. Simple adjustment for E_4 would not be enough. However, regarding the first mechanism dis-cussed above (i.e. C_1 is present for other mechanisms than A_1) analyses could be restricted to the level(s) of E_4 (e.g. $E_4 = 0$) in which the probability of $C_1 = 1$ is the lowest. For the second

mechanism (i.e. interaction between A_1 and E_4), analyses could be restricted to the value(s) of E4 (e.g. $E_4 = 0$) that does not modify the effect of A_1 on C_1, even if the identification of these E_4 strata is not easy and should be based on external information. In addition, it could be possible to focus the analyses directly on the C_1 and C_2 characteristics, instead of using the cancer subtypes as the outcome. Associational estimates would be more inter-pretable, provided that each set of characteristics is ana-lyzed independently, without adjusting for the other characteristics.

It should be noted that the problems described here relate to the underlying biological mechanisms leading to the tumor subtypes, and are therefore different from problems of measurement error and misclassification of the characteristics C_s. They occur even if the characteristics C_s are perfectly measured. It should be also noted that in Fig. 2 we implicitly assumed that the sets of characteristics C_1 and C_2 were independent biomarkers of the pathogenic mechanism A_1, implicitly assuming that C_1 could not cause C_2. It is however possible that early molecular changes affect later changes, so it could be possible to add an arrow from C_1 to C_2, without altering the concepts discussed in this paragraph.

Scenario 2: lack of causality

A different problem may occur if one of the supposed pathogenic mechanisms leading to the characteristics used to define the subtypes is not a cause of the tumor (i.e. the mechanism was wrongly considered as potentially patho-genic for the disease of interest). The DAG shown in Fig. 3 depicts such a scenario in which A_2 is not a mechanism by which the exposures E_2 and E_3 cause the disease, and thus there is not an arrow from A_2 to T. The corresponding characteristic C_2 could still have a prognostic role, but it would not be a marker of a pathogenic mechanism. For example, methylation in the promoter of the O6-methyl-guanine methyltransferase (MGMT) gene in glioblastoma affects response to treatment with telozolemide (and it is thus relevant clinically), but this molecular alteration might not be a driver in the pathogenesis of this cancer type [12]. Under this scenario, any risk factor for A_2 (and thus C_2) would be associated with both disease subtypes defined on the basis of C_2 (presence or absence of C2) even if this risk factor does not have any aetiological role for the tumor of interest. It should be noted that the lack of an arrow from A_2 to T in the DAG shown in Fig. 3 strictly implies no causal effect of A_2 on T. In practice, lack of causality is difficult to prove and molecular changes that may seem to be unrelated with tumor development could eventually be found to have a causal effect. For example, an involvement of MGMT methylation in glioblastoma development is possible, even if the main importance of this marker remains clinical [13].

Numerical examples

The numerical examples presented in this section have an illustrative purpose. It would be thus possible to assess different scenarios or use different values for the various parameters involved in Tables 1 and 2. We chose arbitrary values for the parameters, but we have been careful to avoid implausible or extreme values.

In Table 1, we use the scenario of Fig. 2b and consider a tumor caused by two possible pathogenic mechanisms: A_1 which always causes the characteristics C_1 and C_2, and A_2 which always causes only the characteristic C_2. We also assume that an exposure E_3 doubles the risk of A_2. We assume that the population risk of A_1 is 1 per 1000 and that the population risk of A_2 is 1 per 1000 when $E_3 = 0$ and 2 per 1000 when $E_3 = 1$. We assume complete case ascer-tainment and diagnosis, irrespective of the mechanism involved, and that all variables are measured with no error. We also assume a 10% risk of developing the characteristic C_1 for reasons that are unrelated with the disease of interest (compared to 100% of subjects who are affected by A_1). Finally, we assume that the two mechanisms A_1 and A_2 can coexist, even if in this particular example this assumption has negligible effects because the risks of A_1 and A_2 are low.

Using these assumptions, Table 1 gives an example of the weak effect scenario caused when C_1 can occur also in individuals without the disease of interest. Although the exposure E_3 affects only the pathogenic mechanism A_2 (which corresponds to tumor subtype 2 - S = 2 -), analyses based on tumor subtypes would suggest that E_3 also affects S = 1 (which corresponds to the unaffected pathogenic mechanism A_1).

In Table 2, we also assume that only 50% of the subjects having the pathogenic mechanism A_1 will in fact have the characteristic C_1. This, in combination with the occurrence of C_1 in subjects without the disease, further contributes to the interpretative problems due the weak effect scenario: the risk ratio for S = 2 (corresponding to the pathogenic mechanism A_2) underestimates the effect of E_3 on A_2 (RR of 1.67 vs. a true RR of 2.00), while the risk ratio for S = 1 further overestimates the effect of E_3 on A_1 (RR of 1.15 vs. a true RR of 1.00). Analyses focused on characteristics C_1 and C_2 would reveal no association between E_3 and C_1 and a relative risk of 1.5 of C_2 for E_3 (data not shown in Tables).

Finally, Table 3 gives an example of the lack of causality scenario. There are only two possible subtypes (based on the presence or absence of C_2) as A_2 is not causal for the disease, and thus the pathogenic mechanism A_1 (and $C_1 = 1$) is a necessary cause. In Table 3 we assume that the population risk of A_1 is 1/1000 irrespectively of the level of the exposure E_3, while E_3 doubles the risk (from 10 to 20%) of another mechanism A_2 unrelated with the tumor of interest. We also assume that A_1 always causes C_1, A_2 always causes C_2, no case is left undiagnosed and all variables are measured correctly. Since the characteristic C_2 is unrelated with any pathogenic mechanism for the disease, when the analyses are stratified on C_2, the overall

Table 1 Example of a weak effect scenario

Exposure	Pathogenic mechanism[a]		Subtype[b]	
	A_1	A_2	S = 1 ($C_1 = 1$, $C_2 = 1$)	S = 2 ($C_1 = 0$, $C_2 = 1$)
$E_3 = 1$	0.001	0.002	0.0012	0.0018
$E_3 = 0$	0.001	0.001	0.0011	0.0009
Risk ratio[c]	1.00	2.00	1.09	2.00

Probabilities of different combinations of presence of an exposure E_3, pathogenic mechanisms A_1 and A_2, characteristics C_1 and C_2 and subtypes S = 1 ($C_1 = 1$ and $C_2 = 1$) and S = 2 ($C_1 = 0$ and $C_2 = 1$), when the characteristic C_1 may occur also among subjects without the disease. Relationships among the variables are described in Fig. 2b

[a] These probabilities are assumed by design (see text for details)

[b] These probabilities are obtained on the basis of A_1 and A_2, assuming that A_1 always generates C_1 and A_2 always generates C_2 and a 10% risk of developing the characteristic C_1 for reasons that are unrelated with the disease of interest (see text)

[c] Risk ratios are calculated by dividing the disease probabilities in $E_3 = 1$ for disease probabilities in $E_3 = 0$. Other approaches are possible, which, however, would give the same results as we assumed a rare disease

Table 2 Example of a weak effect scenario

Exposure	Pathogenic mechanism[a]		Subtype[b]	
	A_1	A_2	S = 1 ($C_1 = 1$, $C_2 = 1$)	S = 2 ($C_1 = 0$, $C_2 = 1$)
$E_3 = 1$	0.001	0.002	0.00075	0.00225
$E_3 = 0$	0.001	0.001	0.00065	0.00135
Risk ratio[c]	1.00	2.00	1.15	1.67

Probabilities of different combinations of presence of an exposure E_3, pathogenic mechanisms A_1 and A_2, characteristics C_1 and C_2 and subtypes S_1 ($C_1 = 1$ and $C_2 = 1$) and S_2 ($C_1 = 1$ and $C_2 = 1$), when the characteristic C_1 may occur also among subjects without the disease and the pathogenic mechanism A_1 does not always produce C_1. Relationships among the variables are described in Fig. 2b

[a] These probabilities are assumed by design (see text for details)

[b] These probabilities are obtained on the basis of A_1 and A_2 and assuming that A_1 always causes C_2 and causes C_1 in 50% of the affected subjects, while A_2 always causes C_2. We also assume a 10% probability of C_1 in subjects not having A_1 (see text)

[c] Risk ratios are calculated by dividing the disease probabilities in $E_3 = 1$ for disease probabilities in $E_3 = 0$. Other approaches are possible, which, however, would give the same results as we assumed a rare disease

Table 3 Example of a lack of causality scenario

Exposure	Pathogenic mechanism[a]		Subtype[b]	
	A_1 (causing the disease)	A_2 (not causing the disease)	non C_2-subtype	C_2-subtype
$E_3 = 1$	0.001	0.2	0.0008	0.0002
$E_3 = 0$	0.001	0.1	0.0009	0.0001
Risk ratio[c]	1.00	2.00	0.89	2.00

Probabilities of different combinations of presence of an exposure E_3, pathogenic mechanisms A_1 and A_2 (the latter is not a cause of the disease of interest), characteristics C_1 and C_2, and C_2-subtype ($C_1 = 1$ and $C_2 = 1$) and non C_2-subtype ($C_1 = 1$ and $C_2 = 0$). The relationships among the variables are described in Fig. 3

[a] These probabilities are assumed by design (see text for details)

[b] These probabilities are obtained on the basis of A_1 and A_2 and assuming that only A_1 causes the disease of interest

[c] Risk ratios are calculated by dividing the disease probabilities in $E_3 = 1$ for disease probabilities in $E_3 = 0$. Other approaches are possible, which, however, would give the same results as we assumed a rare disease

group of cases is divided into two groups which are in fact homogeneous in terms of aetiology. An exposure E_3 affecting C_2 would thus increase the number of patients labeled as having a ''C_2 tumor subtype'', and thus decrease the number of patients with a ''non-C_2 like tumor''. It follows that E_3 would be incorrectly considered to be a risk factor for the C_2 tumor subtype and a protective factor for the non-C_2 tumor subtype. Analyses focused on the char-acteristics C_1 and C_2 would reveal no association between E_3 and C_1 and a relative risk of 2.0 of C_2 for E_3 (data not shown in Tables).

Discussion

Categorization of tumors into homogenous subtypes can help to identify specific pathogenic mechanisms and thereby specific risk factors [14]. There has recently been increasing interest in the definition and detection of tumor subtypes. This has been given impetus by recent advances in various omics technologies and the resulting increase in the amount of clinically relevant information that is available. Intertumor heterogeneity [15] and molecular pathological epidemiology [16–18] are fast developing and influential concepts in cancer research. Molecular patho-logical epidemiology in particular has the specific aim of considering molecular pathological signatures to identify more homogenous subtypes in terms of etiology and prognosis [19, 20]. This concept has been used also to discuss, in biological terms, apparently paradoxical asso-ciations that have been largely debated in recent epidemi-ological literature, namely the situation where a disease risk factor is associated with a favorable prognosis among subjects with that disease [21, 22].

A similar increasing attention to the identification of disease subtypes has been occurring for non-cancer dis-eases. For example, already in 1995, an influential paper identified three asthma phenotypes based on combinations of age at onset and persistency of wheezing [23]. This classification, as well as subsequent developments includ-ing additional clinical characteristics and inflammation markers [24] have been used in a large number of aetio-logical studies on the assumption that ''considering these more homogeneous phenotypes in future studies may lead to a better identification of risk factors for asthma'' [25].

Etiological research of cancer subtypes imposes com-plex methodological challenges, which we have not been discussed in this manuscript. First, although epidemiolog-ical research has traditionally focused mainly on multiple exposures, studies of subtypes should use adequate designs and statistical methods to deal with multiple outcomes and to quantify their heterogeneity [26]. Second, for the sake of simplicity, we have considered scenarios in which the characteristics to classify the subtypes are binary but, as recently discussed [20], biomarkers may be ordinal or even continuous, thus leading to a wide spectrum of disease subtypes and challenges in the definitions of the cutoffs. Furthermore, information from several biomarkers can be combined leading to an even higher level of complexity. Finally, measurement of biomarkers, even when binary, is almost inevitably associated with a certain degree of errors and lack of reproducibility. This applies for example to consolidated measurements of ESR1, PGR and ERBB2 by immunohistochemistry in breast cancer [27] and is undoubtedly a strong caveat when several markers are measured through-omics techniques.

As discussed in this paper, the markers used to sub-classify tumors into tumor subtypes should be as close as possible to the actual pathogenic mechanisms. Although this correspondence is crucial to avoid interpretative biases, our knowledge of the pathogenic mechanisms is often too limited to assess whether this important criterion has been achieved. The classification of breast cancer subtypes, for example, based on the expression of the estrogen and progesterone receptors and ERBB2, has been suggested through a cluster analysis and its clinical value has been proven. However, even if studies have suggested that known breast cancer risk factors could have different effects depending on breast cancer molecular subtypes [8, 28], distinct pathogenic mechanisms for the different subtypes have not yet been demonstrated, while it is clear that breast cancer is characterized by a complex molecular heterogeneity [29, 30]. There is thus a tension between the identification of molecular features and the possibility to use them in etiological research. A recent commentary on breast cancer suggests the existence of only two aetiolog-ical components (which would correlate with the expres-sion of the ESR1); it argues that, even if the model may seem too simplistic clinically, it is not too simple for aetiological purposes, considering that many molecular alterations may be more linked with tumor progression than with its development [31].

We suggest that aetiological research into tumor subtypes should first aim to connect the pathogenic mechanisms to the relevant characteristics, and then use these characteristics to assess whether the disease sub-types have different risk factors. Biological knowledge is a key factor. For example, when the subtypes are iden-tified on the basis of the cell type of origin, it can be reasonably assumed, solely on a biological basis, that different cell types are involved in different pathogenic mechanisms. They may share risk factors, as, for example, small cell lung carcinoma and lung adenocar-cinoma are both affected by smoking [32], but the pathogenic mechanisms remain different as they involve different cell types.

Often, however, biological knowledge is not sufficient to link a characteristic to a pathogenic mechanism and research should be conducted with the primary aim of establishing such a link. There are several options. First, characteristics that are evident at an early tumor stage are more likely to be causally linked to its aetiology than late characteristics. Tumor cells evolve during the tumor lifespan, acquiring new and complex molecular features. If we are however inter-ested in primary prevention and early development of the tumor, molecular characteristics acquired at a later stage are less relevant and may easily be affected by mechanisms that are not related with the risk factors of interest. Thus, studies that have access to pre-diagnostic tissues are highly infor-mative to define tumor subtypes for aetiological studies [33]. Once the subtypes are defined, they can be identified also on the diagnostic tissue, but an initial step involving pre-diag-nostic tissue and early molecular characteristics would greatly enhance the potential to validly interpret subsequent studies. Second, the risk of interpretative bias may be reduced by defining tumor subtypes on the basis of subtype-specific sets of characteristics (i.e. each subtype has different identifying characteristics) instead of combinations of characteristics partially shared by different pathogenic mechanisms. This should be taken into account, for example, when the tumor subtypes are defined on the basis of an unsupervised cluster analysis, and then a set of markers is chosen to characterize each specific cluster. For aetiological research, it is perhaps safer if the characterizing sets of markers do not overlap among clusters. Third, in some instances it is possible to directly test whether a characteristic is causally involved in tumor development. For example, to understand whether gene-specific methylation is causally involved in tumor development, it is possible to study the association between germ-line variants in the DNA methy-lation machinery genes and cancer incidence [34]. If an association is found, methylation markers are more likely to be causally involved instead of being just epiphenomena. This approach, which is based on the concept of Mendelian randomization, can be carried out for characteristics that are known to be affected by germ-line variants.

In conclusion, categorization of tumors into homoge-nous subtypes may have important implications for aetio-logical research and identification of risk factors. However, it is essential that the characteristics used to classify tumors into subtypes should be as close as possible to the actual pathogenic mechanisms to avoid interpretative biases. Whenever our knowledge of these mechanisms is limited, research into risk factors for tumor subtypes should first aim to causally link the characteristics to the pathogenic mechanisms.

Acknowledgements We would like to thank Dr. Andreas Pettersson for helpful comments on earlier versions of this paper.
Funding Lorenzo Richiardi was partially supported by a Fulbright Research Scholar fellowship when working on this paper. The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013)/ERC Grant Agreement No. 668954.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Collins FS, Varmus H. A new initiative on precision medicine. N Engl J Med. 2015;372(9):793–5.
2. Muraro A, Lemanske RF Jr, Hellings PW, Akdis CA, Bieber T, Casale TB, et al. Precision medicine in patients with allergic diseases: airway diseases and atopic dermatitis-PRACTALL document of the European Academy of Allergy and Clinical Immunology and the American Academy of Allergy, Asthma and Immunology. J Allergy Clin Immunol. 2016;137(5):1347–58.
3. Pitt GS. Cardiovascular precision medicine: hope or hype? Eur Heart J. 2015;36(29):1842–3.
4. Pearson ER. Personalized medicine in diabetes: the role of 'omics' and biomarkers. Diabet Med. 2016;33(6):712–7.
5. Blows FM, Driver KE, Schmidt MK, Broeks A, van Leeuwen FE, Wesseling J, et al. Subtyping of breast cancer by immunohisto-chemistry to investigate a relationship between subtype and short and long term survival: a collaborative analysis of data for 10,159 cases from 12 studies. PLoS Med. 2010;7(5):e1000279.
6. Network Cancer Genome Atlas. Comprehensive molecular por-traits of human breast tumours. Nature. 2012;490(7418):61–70.
7. Goldhirsch A, Winer EP, Coates AS, Gelber RD, Piccart-Gebhart M, Thu¨rlimann B, et al. Personalizing the treatment of women with early breast cancer: highlights of the St Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer 2013. Ann Oncol. 2013;24(9):2206–23.
8. Yang XR, Chang-Claude J, Goode EL, Couch FJ, Nevanlinna H, Milne RL, et al. Associations of breast cancer risk factors with tumor subtypes: a pooled analysis from the Breast Cancer Association Consortium studies. J Natl Cancer Inst. 2011;103(3): 250–63.
9. WHO/IARC Classification of tumours, http://publications.iarc.fr/ Book-And-Report-Series/Who-Iarc-Classification-Of-Tumours. Accessed 28 Jul 2016.
10. Song Q, Merajver SD, Li JZ. Cancer classification in the genomic era: five contemporary problems. Hum Genom. 2015;9:27.
11. Blows FM, Driver KE, Schmidt MK, Broeks A, van Leeuwen FE, Wesseling J, et al. Subtyping of breast cancer by immunohisto-chemistry to investigate a relationship between subtype and short and long term survival: a collaborative analysis of data for 10,159 cases from 12 studies. PLoS Med. 2010;7(5):e1000279.
12. Hegi ME, Diserens AC, Gorlia T, Hamou MF, de Tribolet N, Weller M, et al. MGMT gene silencing and benefit from temo-zolomide in glioblastoma. N Engl J Med. 2005;352(10): 997–1003.
13. Rapkins RW, Wang F, Nguyen HN, Cloughesy TF, Lai A, Ha W, et al. The MGMT promoter SNP rs16906252 is a risk factor for MGMT methylation in glioblastoma and is predictive of response to temozolomide. Neuro Oncol. 2015;17(12):1589–98.
14. Morton LM, Slager SL, Cerhan JR, Wang SS, Vajdic CM, Ski-bola CF, et al. Etiologic heterogeneity among non-Hodgkin lymphoma subtypes: the InterLymph non-Hodgkin lymphoma

subtypes project. J Natl Cancer Inst Monogr. 2014;2014(48): 130–44.
15. Zellmer VR, Zhang S. Evolving concepts of tumor heterogeneity. Cell Biosci. 2014;4:69.
16. Ogino S, Stampfer M. Lifestyle factors and microsatellite insta-bility in colorectal cancer: the evolving field of molecular pathological epidemiology. J Natl Cancer Inst. 2010;102(6): 365–7.
17. Ogino S, Lochhead P, Chan AT, Nishihara R, Cho E, Wolpin BM, et al. Molecular pathological epidemiology of epigenetics: emerging integrative science to analyze environment, host, and disease. Mod Pathol. 2013;26(4):465–84.
18. Ikram MA. Molecular pathological epidemiology: the role of epidemiology in the omics-era. Eur J Epidemiol. 2015;30(10): 1077–8.
19. Hamada T, Keum N, Nishihara R, Ogino S. Molecular patho-logical epidemiology: new developing frontiers of big data sci-ence to study etiologies and pathogenesis. J Gastroenterol. 2016 Oct 13.
20. Ogino S, Nishihara R, VanderWeele TJ, Wang M, Nishi A, Lochhead P. Review article: the role of molecular pathological epidemiology in the study of neoplastic and non-neoplastic dis-eases in the era of precision medicine. Epidemiology. 2016;27(4):602–11.
21. Nishihara R, VanderWeele TJ, Shibuya K, Mittleman MA, Wang M, Field AE, et al. Molecular pathological epidemiology gives clues to paradoxical findings. Eur J Epidemiol. 2015;30(10): 1129–35.
22. Porta M, Vineis P, Bolu´mar F. The current deconstruction of paradoxes: one sign of the ongoing methodological ‘‘revolution’’. Eur J Epidemiol. 2015;30(10):1079–87.
23. Martinez FD, Wright AL, Taussig LM, Holberg CJ, Halonen M, Morgan WJ. Asthma and wheezing in the first six years of life. The Group Health Medical Associates. N Engl J Med. 1995;332(3):133–8.
24. Wenzel SE. Asthma phenotypes: the evolution from clinical to molecular approaches. Nat Med. 2012;185:716–25.
25. Siroux V, Basagan˜a X, Boudier A, Pin I, Garcia-Aymerich J, Vesin A, et al. Identifying adult asthma phenotypes using a clustering approach. Eur Respir J. 2011;38(2):310–7.
26. Wang M, Spiegelman D, Kuchiba A, Lochhead P, Kim S, Chan AT, et al. Statistical methods for studying disease subtype heterogeneity. Stat Med. 2016;35(5):782–800.
27. Laible M, Schlombs K, Kaiser K, Veltrup E, Herlein S, Lakis S, Sto¨hr R, Eidt S, Hartmann A, Wirtz RM, Sahin U. Technical validation of an RT-qPCR in vitro diagnostic test system for the determination of breast cancer molecular subtypes by quantifi-cation of ERBB2, ESR1, PGR and MKI67 mRNA levels from formalin-fixed paraffin-embedded breast tumor specimens. BMC Cancer. 2016;7(16):398.
28. Tamimi RM, Colditz GA, Hazra A, Baer HJ, Hankinson SE, Rosner B, et al. Traditional breast cancer risk factors in relation to molecular subtypes of breast cancer. Breast Cancer Res Treat. 2012;131(1):159–67.
29. Heng YJ, Lester SC, Tse GM, Factor RE, Allison KH, Collins LC, et al. The molecular basis of breast cancer pathological phenotypes. J Pathol. 2017;241(3):375–91.
30. Denkert C, Liedtke C, Tutt A, von Minckwitz G. Molecular alter-ations in triple-negative breast cancer-the road to new treatment strategies. Lancet. 2016. doi:10.1016/S0140-6736(16)32454-0.
31. Anderson WF, Rosenberg PS, Prat A, Perou CM, Sherman ME. How many etiological subtypes of breast cancer: two, three, four, or more? J Natl Cancer Inst. 2014;106(8):dju165.
32. Pesch B, Kendzia B, Gustavsson P, Jo¨ckel KH, Johnen G, Pohla-beln H, et al. Cigarette smoking and lung cancer–relative risk estimates for the major histological types from a pooled analysis of case-control studies. Int J Cancer. 2012;131(5):1210–9.
33. Kadara H, Scheet P, Wistuba II, Spira AE. Early events in the molecular pathogenesis of lung cancer. Cancer Prev Res. 2016;9(7):518–27.
34. Sung H, Yang HH, Zhang H, Yang Q, Hu N, Tang ZZ, et al. Common genetic variants in epigenetic machinery genes and risk of upper gastrointestinal cancers. Int J Epidemiol. 2015;44(4): 1341–52.