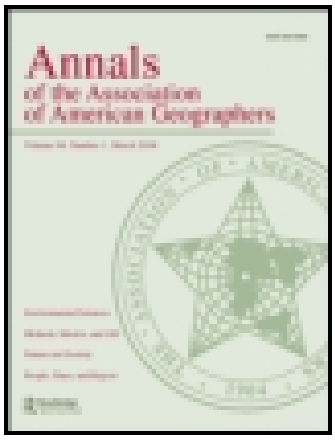

Annals of the Association of American Geographers
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/raag20

Genetic GIScience: Toward a Place-Based Synthesis of the Genome, Exposome, and Behavome

Geoffrey M. Jacquez[a], Clive E. Sabel[b] & Chen Shi[c]

[a] Department of Geography, University at Buffalo—State University of New York, and BioMedware

[b] School of Geographical Sciences, University of Bristol

[c] Department of Geography, University at Buffalo—State University of New York

Published online: 27 Apr 2015.

Genetic GIScience: Toward a Place-Based Synthesis of the Genome, Exposome, and Behavome

Geoffrey M. Jacquez,* Clive E. Sabel,† and Chen Shi‡

*Department of Geography, University at Buffalo—State University of New York, and BioMedware
†School of Geographical Sciences, University of Bristol
‡Department of Geography, University at Buffalo—State University of New York

The *exposome*, defined as the totality of an individual's exposures over the life course, is a seminal concept in the environmental health sciences. Although inherently geographic, the exposome as yet is unfamiliar to many geographers. This article proposes a place-based synthesis, genetic geographic information science (genetic GIScience), that is founded on the exposome, genome+, and behavome. It provides an improved understanding of human health in relation to biology (the genome+), environmental exposures (the exposome), and their social, societal, and behavioral determinants (the behavome). Genetic GIScience poses three key needs: first, a mathematical foundation for emergent theory; second, process-based models that bridge biological and geographic scales; third, biologically plausible estimates of space–time disease lags. Compartmental models are a possible solution; this article develops two models using pancreatic cancer as an exemplar. The first models carcinogenesis based on the cascade of mutations and cellular changes that lead to metastatic cancer. The second models cancer stages by diagnostic criteria. These provide empirical estimates of the distribution of latencies in cellular states and disease stages, and maps of the burden of yet to be diagnosed disease. This approach links our emerging knowledge of genomics to cancer progression at the cellular level, to individuals and their cancer stage at diagnosis, and to geographic distributions of cancer in extant populations. These methodological developments and exemplar provide the basis for a new synthesis in health geography: genetic GIScience. *Key Words: cancer epidemiology, dynamical systems, genetic GIScience, health geography, space–time.*

环境暴露（exposome），定义为个人一生中的总体暴露量，是环境健康科学中一个具有影响力的概念。儘管环境暴露在本质上是地理的，但却尚未被诸多地理学者所熟知。本文提出一个根据地方的综合性基因地理信息科学（genetic GIScience），该科学以环境暴露、基因组+与环境暴露的决定因素（behavome）为基础。此一概念，提供了人类健康之于生物（基因组+）、环境暴露（exposome），及其社会互动、社会组织结构和行为的决定因素（behavome）的改良式理解。基因地理信息科学，提出三大主要需求：首先，是新兴理论的数学基础；再者，是连结生物与地理尺度、以过程为基础的模型；第三，是对空间—时间的传染病迟滞进行生物学的可靠估计。划分的模型，是一个可能的解决方法；本文运用胰脏癌作为范例，建立两种模型。第一个模型，根据导致转移型癌症的一系列变异与细胞改变，将致癌作用进行模式化。第二个模型，则根据诊断标准，模式化癌症阶段。这些模型，提供了细胞状态与疾病阶段中，潜伏因素的分佈之经验性估计，以及尚未被诊断出的疾病之压力描绘。本研究方法，将我们正在形成的基因知识，连结至细胞层级的癌症进程、个人及其确诊时的癌症阶段，以及癌症在现今人口中的地理分佈。这些方法论的发展与范例，为健康地理学提供了一个崭新的综合体之基础：基因地理信息科学。 *关键词：癌症流行病学，动态系统，基因地理信息科学，健康地理学，空间—时间。*

El exposoma, definido como la totalidad de las exposiciones a las que se somete un individuo en el curso de la vida, es un concepto seminal en las ciencias de la salud ambiental. Aunque es inherentemente geográfico, hasta ahora el exposoma no es familiar para la mayoría de los geógrafos. Este artículo propone una síntesis basada en lugar, la ciencia de la información geográfica genética (SIGciencia genética), la cual está fundamentada en el exposoma, el genoma+ y el "conductoma" (*behavome*). Tal síntesis facilita un mejor entendimiento de la salud humana en relación con la biología (el genoma+), las exposiciones ambientales (el exposoma) y sus determinantes sociales, sociológicos y conductuales (el conductoma). La SIGciencia genética plantea tres necesidades claves: primera, una fundamentación matemática para la teoría emergente; segunda, modelos basados en proceso que conecten las escalas biológica y geográfica; tercera, estimativos biológicamente plausibles de los rezagos espacio-temporales de la enfermedad. Los modelos compartimentados pueden ser una posible solución; este artículo desarrolla dos modelos, utilizando como referente el cáncer pancreático. El primero modela la carcinogénesis con base en la cascada de mutaciones y cambios celulares que desembocan en cáncer metastásico. El segundo modela las etapas del cáncer con

Annals of the Association of American Geographers, 0(0) 2015, pp. 1–19 © 2015 by Association of American Geographers
Initial submission, August 2013; revised submissions, February and May 2014; final acceptance, May 2014
Published by Taylor & Francis, LLC.

criterios de diagnóstico. Estos modelos proveen estimativos empíricos de la distribución de latencias en estados celulares y etapas de la enfermedad, y mapas de la carga de una enfermedad que todavía está por ser diagnosticada. Tal enfoque enlaza nuestro emergente conocimiento de la ciencia del genoma con la progresión del cáncer a nivel celular, con los individuos y su etapa del cáncer en diagnóstico, y con las distribuciones geográficas del cáncer en poblaciones existentes. Estos desarrollos metodológicos y el ejemplo son la base para una nueva síntesis en geografía de la salud: la SIGciencia genética. *Palabras clave: epidemiología del cáncer, sistemas dinámicos, SIGciencia genética, geografía de la salud, espacio-tiempo.*

Environment, social, and individual factors all play a role in an individual's health and well-being. Linking social and health data to a particular location is important because where we live can and does influence our health (Tunstall, Shaw, and Dorling 2004). Health outcomes are related to an individual's physical and social environment, including factors such as water, soil, and air content; exposure to hazardous materials; tobacco smoke; occupation; marital status; social support; and characteristics of the home, in addition to the composition of the local built environment (Marmot 2000; Pickle, Waller, and Lawson 2005).

Geographical epidemiology rests largely on the assumption that the spatial incidence of diseases holds a key to their cause. High population mobility, long latent periods, and environmental change complicate matters, however, distorting what might otherwise be a direct relationship between cause and effect (G. M. Jacquez 2004; Kwan 2009). This gives rise to what has been called the *space–time lag* (Dearwent, Jacobs, and Halbert 2001; Griffith and Paelinck 2009). From a geographical point of view, this means that the place or environment where the case is discovered and diagnosed is not necessarily the same place or environment where the exposure occurred (Picheral 1982; Sabel et al. 2000; Sabel et al. 2003).

Many studies examining associations between geographical patterns of health and disease and causal factors assume that current residence in an area can be equated with exposure to conditions that currently (and historically) pertain there (Bentham 1988). This is important, as the place of residence at the time of diagnosis or death is often adopted by epidemiologists and geographers as the location for further analysis of the disease in question. Yet people move, and hence previous exposure to pathogens will not be included in the study. The problems will be greater for diseases that have a long lag or latency period, allowing plenty of time for mobility of the population. By adopting only the current residential address, not only will an individual's migration history be neglected but additionally the daily "activity spaces" and associated uncertainties will be ignored (G. M. Jacquez 2004; Kwan 2012). For chronic diseases such as cancer, we often use the place of residence at diagnosis or death to record the health event. But where people reside at time of diagnosis could be far removed from where they lived when causative exposures occurred. This disconnect is widely recognized (Wheeler, Ward, and Waller 2012), yet techniques for estimating appropriate sampling distributions for latencies for the geographic modeling of human diseases that are biologically reasonable, based on observable disease states, and that incorporate knowledge of disease progression are seldom available. This article attempts to address this need using the construct of genetic GIScience.

Space–Time Geovisualization and Modeling

There is a long and rich history of geographers investigating space–time interaction from Hägerstraand through Forer to the present (Hägerstraand 1970; Forer 1978; Richardson et al. 2013). Interest has focused on the space–time cube. Space–time paths or geospatial lifelines have largely been used to visualize (often neglecting modeling) individual mobilities through space, such as Forer's work in Auckland visualizing student lifestyles (Huisman and Forer 1998). Kwan (2000) has used the cube—she uses the term *space–time aquarium*—to study accessibility differences among gender and different ethnic groups in Portland. Miller (1999) applied its principles in trying to establish accessibility measures in an urban environment. For physical environmental exposures, Hedley et al. (1999) created an application in a geographical information system (GIS) environment for radiological hazard exposure, and Gulliver and Briggs (2005) modeled space–time interactions to traffic derived air pollution. Others have assessed similarity in geospatial lifelines and clustered them to quantify disease patterns for mobile populations (Sinha and Mark 2005; G. M. Jacquez et al. 2013). Modeling latency between exposure and disease outcomes largely remains neglected, however.

Recent improved data gathering techniques, including the wider availability of Global Positioning System (GPS), cell phone, and social media data have renewed interest in time geography and the space–time cube. Dykes and Mountain (2003) discussed data collection techniques by mobile phone, GPS, and location-based services and suggested a visual analytical method to deal with the data gathered. Lam (2012) and Bian et al. (2012) both discussed ongoing challenges to health risk assessment despite the wider availability of individual-level data.

The Exposome, Genome+, and Behavome

A new synthesis in health geography we are calling genetic GIScience seeks to document, quantify, and model the relationships among place, the genome, the exposome (Wild 2005, 2012), and the behavome that are the determinants of illness and wellness (Figure 1). This builds on and extends prior constructs in human biology and ecology, such as the nature–nurture debate and Meade's triangle of human ecology, which viewed health outcomes as the result of place- and time-specific interactions among populations, their environments, and their behaviors (Meade 1977).

This paradigm requires an explicit understanding of how these determinants are related to space–time patterns of health outcomes in human populations. Because the genome, exposome, and behavome are defined at the level of the individual, techniques for estimating disease latency—the time between the onset of disease and its diagnosis—are essential. A second key need is the ability to integrate and model the influence of the genome, exposome, and behavome across biological scales and to then geographically map the results at local and regional scales. Finally, sound theory often requires a solid mathematical foundation, and one must be established for genetic GIScience. This article seeks to begin to mathematically formalize these requirements, and poses an example using one of the least understood cancers, pancreatic cancer.

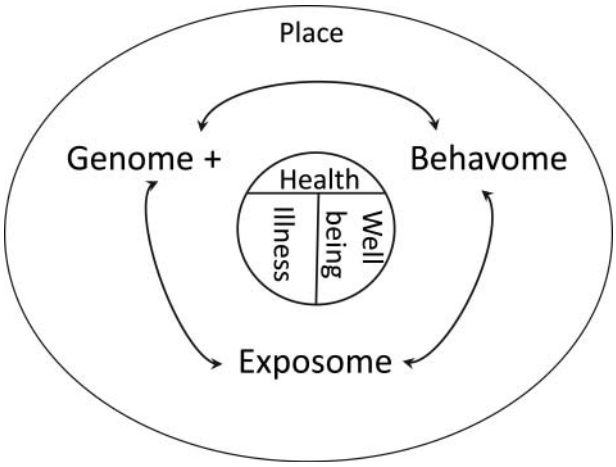

Figure 1. Schematic representation of genetic geographic information science (genetic GIScience). The three primary determinants of health, both in terms of illness and well-being, are (1) an individual's biology, which can quantified as his or her genome+, made up of the genome (genetic composition), regulome (which controls gene expression), proteome (complement of amino acids and proteins), and metabalome (the basis of metabolism and homeostasis); (2) the environments he or she experiences, which could be quantified as the exposome, defined as the totality of exposures over the life course (Wild 2005); (3) the totality of an individual's health behaviors over the life course, which could be quantified as the behavome and mediate the exposome and interactions between the exposome and the genome+. These determinants of human health act through place, defined as the geographic, environmental, social, and societal milieus experienced over a person's life course. This synthesis is referred to as genetic GIScience.

Exposome

The term *exposome* was introduced by Wild (2005) to "encompass life-course environmental exposures (including lifestyle factors), from the prenatal period onwards ... the exposome is a highly variable entity that evolves throughout the lifetime of the individual" (1848). The exposome concept recognizes three broad categories of nongenetic exposures: internal (e.g., metabolism), specific external (e.g., air pollution), and general external (e.g., socioeconomic factors; Wild 2012). Although a person's genome is fixed at conception, internal and external sources of exposure cause the human's internal chemical environment to vary throughout life (Rappaport 2011; Miller and Jones 2014). Essentially, an individual will have a particular profile of exposure at any given point in time that makes the characterization of the exposome so challenging (Wild 2012). It is a concept to measure effects of a lifelong exposure to environmental influences on human health and therefore requires longitudinal sampling especially during fetal development, early childhood, puberty, and the reproductive years (Rappaport 2011). These measures include external monitoring and modeling of media such as air and water but also biomonitoring (i.e., measurements) of biological markers of exposure through methods such as blood or

urine sampling (Lioy and Rappaport 2011). Rappaport (2011) prioritized a top-down approach, applying biomonitoring to identify all important exposures, over a bottom-up approach that is based on air, water, or soil samples to identify all exogenous exposures.

Van Tongeren and Cherrie (2012), on the other hand, supported the aim of developing an integrated concept of exposomics, taking all sources of available exposure information into account. Internal and external exposure data, personal behavior, and environmental measurements could thus be used to determine the exposome. This requires the collaboration of researchers from a variety of disciplines to promote the concept and unravel complex relationships among social interactions, biological effects, and the risk of diseases (Wild 2012), an endeavor suited to but largely unexploited by geographers.

The exposome has a public health-oriented objective and the aim of its application is to aggregate up from a group of individuals to a population, providing the basis for public health decisions (Wild 2012).

Genome+

The genome+ is made up of the individual's genome (genetic composition), regulome (which controls gene expression), proteome (their complement of amino acids and proteins), and metabalome (the basis of metabolism and homeostasis). Together, these constitute a good portion of an individual's biological makeup. The last few years have seen major advances in our ability to quantify the genome+. Technology improvements have dramatically reduced genome sequencing costs. In 2000 the Human Genome Project sequenced the first whole human genome, at a cost of over US$2 billion (Davies 2010). In 2012, the 1000 Genomes Project released their Phase 1 sequencing data (Pybus et al. 2014), documenting genetic variation in more than 1,000 individuals from twenty-five populations from around the globe (The 1000 Genomes Project Consortium 2012). Genome sequencing costs continue to drop, and the US$1,000 whole sequence genome is now available (Hayden 2014). In medical practice and research, whole genome sequencing poses ethical challenges regarding information disclosure to the individual, especially given incomplete knowledge of the genetic basis of disease (Yu et al. 2013). Nonetheless, whole genome sequencing as a commodity will soon cost US$100. Dramatic cost reductions are occurring in the exome, epigenome, and other genome+ constituents (Mefford 2012; Meissner 2012; Weinhold 2012; Zentner and Henikoff 2012). It is clear that measurements of the genome+ will soon be widely and inexpensively available and will be incorporated into individual electronic health records, notwithstanding the informatics and ethical challenges posed by their integration (Hazin et al. 2013; Kho et al. 2013; Tarczy-Hornoch et al. 2013; Flintoft 2014).

As our understanding of the genetic bases of disease has grown, the need for a *systems biology* approach that integrates across genetic, cellular, organ, individual, and population-level scales is increasingly recognized (Orešič 2014). How can we incorporate knowledge, for example, of the cascade of genetic mutations leading to pancreatic cancer into our understanding of cancer latency, and how might this impact estimates of the burden of cancer at the population level? How do changes manifested in pancreatic cells as a result of mutations translate into cancer progression, and can we construct models that capture biological nuance yet are suited to GIScience? For geographers, how can systems biology approaches be integrated into space–time geographic disease models? This article addresses these needs by linking a model of carcinogenesis at the cellular level with a model of cancer stages at the individual and population level.

Behavome

The behavome is made up of an individual's health-related behaviors over the life course and is the most inchoate of the genetic GIScience triad of the genome+, exposome, and behavome. Recognition methods for assessing individual behaviors have been an important research topic for decades. With the advent of sensors in residences, health care facilities, and even wearable sensors on patients, the issue of multisensor data fusion for activity recognition has emerged. These technologies are already being deployed and assessed in nursing home and assisted living facilities but as yet have little penetration in the geographic literature. Recent research has demonstrated that these methods can identify risky behaviors with good accuracy and low deployment costs (Palumbo et al. 2013). The Internet of things, including smart homes, smart cars, and smart workplaces, is in the early phase of what many predict to be explosive growth (Ashton 2009). In 2008 the number of devices on the Internet exceeded the number of people, and in 2020 it will exceed 50 billion devices (Swan 2012).

Information on when, where, and how we use appliances, electronic devices, machinery, and environmental controls in home and workplace settings and while commuting has yet to be used to quantify the behavome. The value of near real-time data on ambient temperatures and how often and when we use the refrigerator might have enormous value for quantifying, for example, personal energy budgets, a key problem in cancer etiology (Ballard-Barbash et al. 2013; Hursting 2014). A variety of different approaches for assessing health behaviors have been suggested using technologies such as inertial sensors, GPS, smart homes, radio frequency identification, and others. Most promising is the sensor fusion approach that combines data from several sensors simultaneously (Lowe and ÓLaighin 2014). To our knowledge, technologies such as Google Glass have yet to be used for capturing video images to chronicle dietary intake and other health-related activities. Other potential applications include quantification of personalized environmental metrics such as individual walkability (e.g., Mayne et al. 2013). Once health-related behaviors are known, the possibility of using gamification (Whitson 2013) and other approaches to encourage salubrious behaviors becomes possible (Schoech et al. 2013).

Contribution of This Article

This article proposes a synthesis of the genome+, exposome, and behavome that is place based and offers a promising new landscape for research in health geography—genetic GIScience. The potential research contributions this synthesis offers geographers are manifold, including health geography, quantitative methods, behavioral geography, visualization, space–time modeling, and social geography. The exposome and behavome are new concepts with many unsolved gaps of their own, several of which are addressed in this article. First, we demonstrate how disease latencies could be estimated using compartmental models and data available from systems biology. Disease latency estimation is a key problem for space–time lags in health geography. Second, space–time models that account for individual disease processes yet provide geographic estimates of disease burden are almost entirely lacking in health geography, a significant gap addressed by this article. Finally, as a motivating example, we develop and apply a comprehensive modeling approach that estimates cancer latency, couples carcinogenesis and stage models, and represents and links processes at the genomic level (e.g., mutation events, cascades of genetic changes that lead to cancer), cellular level (e.g., cell replication and death, DNA repair), organ level (e.g., carcinogenesis in situ and metastases to distant organs), individual level (e.g., cancer staging in the individual, progression of individuals through cancer stages), to the population level (e.g., predicted geographic distributions of undiagnosed cancers). This by no means addresses all of the challenges and gaps posed by genetic GIScience, but it hopefully illustrates the promise of this research direction and perhaps points the way forward.

It is important to note that the breadth of the concept and challenge represented in Figure 1 is substantial. This aim of this contribution is to communicate its scope, identify key research problems, and propose a way forward. The example of pancreatic cancer presented here deals primarily with the genome+. At the time of this writing, measurement of the exposome is at a nascent stage, and the term *behavome* is new. When data from the exposome and behavome become available, they can be incorporated into the modeling framework through place- and person-specific effects on model flow parameters; for example, DNA mutation and repair rates. Opportunities for such adaptation and extension of the modeling framework are identified in the discussion.

We begin with an introduction to the approach for the modeling and analysis of dynamic geographic systems using process-based temporal lags. This is followed by a brief background on latency estimation approaches that motivate the use of residence times in compartmental systems. A primer on compartmental analysis is presented, followed by a simple three-stage model of disease and results for distributions of residence times. Next, the specific example of pancreatic cancer is considered, and a five-state model of carcinogenesis is developed along with its biological foundation. A second model of progression through cancer stages based on diagnostic criteria used by the American Cancer Association follows. These models are linked using knowledge of the mapping of stage of diagnosis with progression of tumor growth and metastatic capacity. This is applied to data from the Michigan cancer registry on stage at diagnosis for all incident pancreatic cancers in white males from 1985 to 2005 in the Detroit metropolitan area. Potential applications of this approach and next steps are then discussed.

The generalized approach (Figure 2, left) applies to any geographic system amenable to a compartmental representation. Here the emphasis is on the

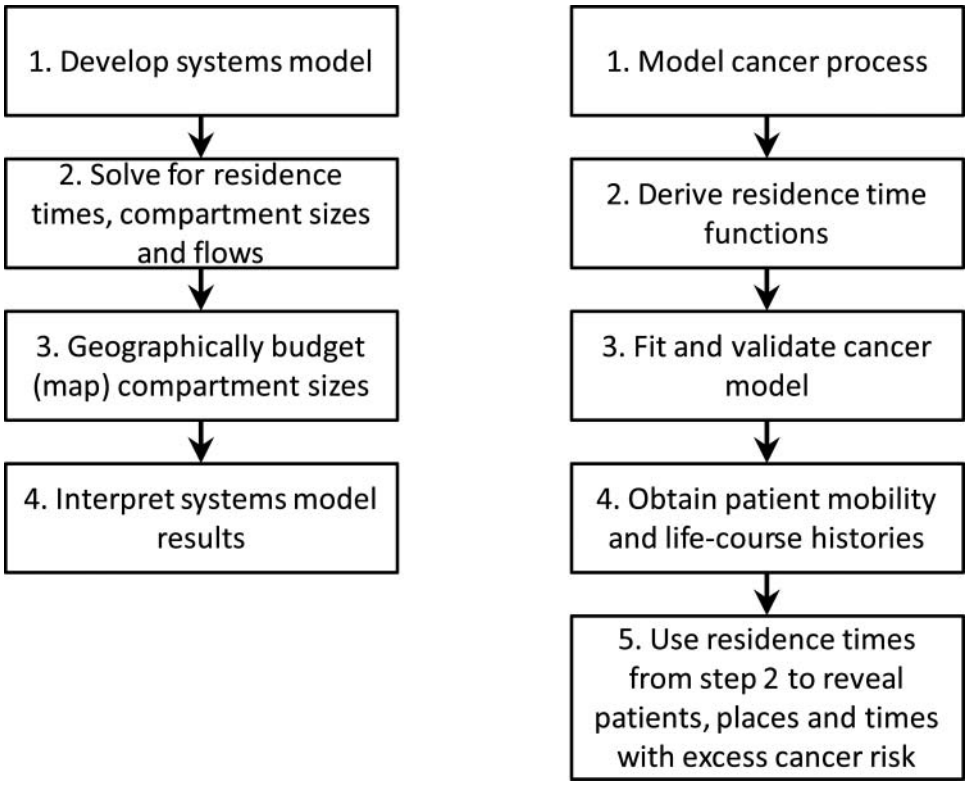

Figure 2. Steps in dynamic geographic systems analysis (left) and specific application to cancer using knowledge of residential history to budget excess risk (right).

development of a minimally sufficient but mechanistically reasonable systems model. An example application to a dynamic geographic systems model of cancer using residence times to estimate latency is shown in Figure 2 (right).

Ideally, genetic GIScience will be based on model-based theory with several characteristics. First, models must be biologically reasonable and capture relevant aspects of disease etiology and natural history. Second, they must provide estimates of the distributions of disease latency. Third, they must be estimable from empirical data, so we can derive latency distributions from observable measures and based on the current state of knowledge of the disease. Fourth, they must provide for geographically referenced data on individuals. The derivation of biologically based estimates of disease latency is a difficult problem, and we next consider alternative approaches to latency estimation.

Disease Latency Estimation

Several techniques exist for modeling disease latency, including representations of cohort exposures, developmental stages of vulnerability, models of empirical induction periods, and compartmental models. We summarize these before focusing on residence times in compartmental systems.

Cohort exposures arise when a common exposure occurs for a group of individuals, resulting in an overall increase in disease risk. Here the temporal lag between the causal event and health outcomes is directly observable. For example, Chernobyl released radioactive iodide over Belarus and led to an increase in pediatric thyroid cancers. The latent period for tumor development was four to six years (Nikiforov and Gnepp 1994).

Developmental stages of vulnerability arise when the timing and characteristics of biological stages of development are associated with increased risk of an adverse health event in later years. For example, genetic risk accounts for approximately 10 to 15 percent of breast cancer cases, and the windows of vulnerability occur before a woman's first birth and during the development of breast tissues (Colditz and Frazier 1995). Here an average latency and its distribution could be estimated as the time from the developmental stage to disease diagnosis.

The *empirical induction period* models latency as the sum of induction and latent periods defined as the periods between causal action and disease initiation (induction) and between disease initiation and detection (latent). The sum of the induction and latent periods is the empirical induction period. The induction period is not estimable except in relation to specific etiologic factors, as different exposures have different levels of effect on disease expression (Rothman 1981).

Residence times in compartmental models of disease can be obtained directly from the model itself. For a given compartmental model and parameter values, the mean residence time and distribution of residence times in each compartment are known. This result holds for both deterministic and stochastic compartmental models but has yet to be used in geographic models of human disease. When the compartments correspond to stages of disease, the compartmental residence times are estimates of disease latency. Compartmental models thus are best constructed so compartments correspond to known disease states (e.g., are biologically reasonable), and the coefficients governing transitions between compartments are formulated in terms of known biological and infection processes (e.g., the mechanics are process-based). Residence times from compartmental models thus convey the characteristics required at the beginning of this section: (1) They can be formulated in a biologically reasonable fashion; (2) they provide estimates of the distribution of latencies; and (3) whether a given model, and hence its residence times, is estimable is known once the model and observable measures are identified. The remainder of this article employs compartmental models. A primer on compartmental models is provided as supplemental material on the publisher's Web site.

Residence Times

Residence times are the time required for a particle to enter and then exit a compartment. Compartment residence times could be used in model validation by comparing residence times from the model to the observed residence times. For linear compartmental models the residence times are inverse exponential functions, and closed-form solutions for calculating the probability density functions are known (J. A. Jacquez 1996). In deterministic nonlinear compartmental systems, the distributions of residence times are functions of the state variables and hence of the compartments' sizes. The probability density functions of linear stochastic models are the same as for their deterministic analog. The probability density functions of residence times for nonlinear stochastic systems differ from those of their nonlinear deterministic counterpart, however (J. A. Jacquez 2002). This article presents residence times linear deterministic stage-based models of cancer, which also apply to their linear stochastic counterparts.

Simplicity versus Complexity, and Implications for Residence Times

There is a tension between simplicity, which makes models easier to understand and mathematically tractable, and complexity, which seeks to incorporate the nuances and details of a complex reality. Simplicity might correspond to a representation with fewer compartments, an implicit combining of compartments that has implications for the modeling of residence times. When the residence times for a compartment in a model are too short, the creation of subcompartments to represent that compartment can be used to obtain longer average residence times (J. A. Jacquez and Simon 2002). Correspondence of residence times to those observed in the system under scrutiny thus can be used as a diagnostic for model oversimplification and misspecification.

Modeling Carcinogenesis: Cancer in the Individual

For the geographic modeling of disease we are interested in identifying places and subpopulations characterized by an excess of cancer for individuals in those states of carcinogenesis when exposures to mutagens might have been causal. That is, we are looking for the geographic signature of the actions of past environmental exposures that gave rise to cancers. To do this we require biologically reasonable models of carcinogenesis (e.g., the biological events that have cancer as their sequelae) and cancer stages (how cancers progress once they have started). We begin with carcinogenesis.

The initial biological event leading to cancer is damage to DNA. Such damage occurs on one DNA strand, and repair mechanisms can reverse that damage. Whether the damage is maintained among daughter cells depends on the timing of replication and

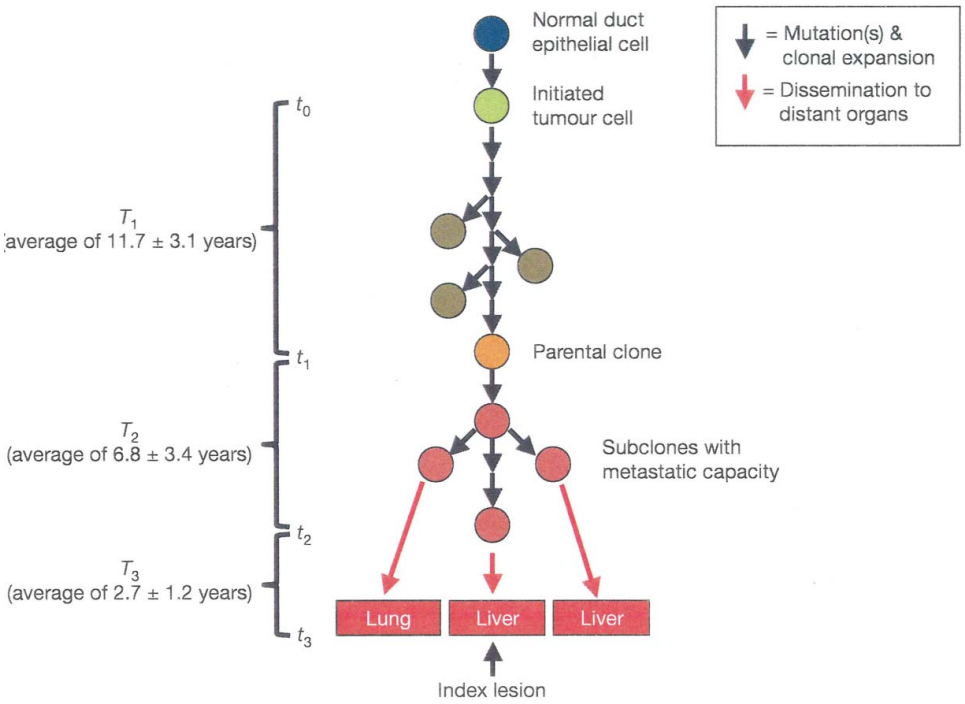

Figure 3. Schematic of the evolution of pancreatic cancer. Normal pancreatic duct epithelial cells undergo mutation events to become an initiated tumor cell. Additional mutations and clonal expansions lead eventually to a founder cell of the index pancreatic cancer clone. These produce subclones with metastatic capacity, eventually leading to dissemination to distant organs such as the liver. Times shown are the empirical residence times in each system state. *Source:* Adapted from Yachida et al. (2010). Reprinted by permission from Macmillan Publishers Ltd.: *Nature*. (Color figure available online.)

repair. If replication occurs before repair, the damaged DNA strand is passed on to the daughter cells (a fixed mutation). Notice that only some of these mutations are deleterious and lead to cancers.

Carcinogenesis models usually treat irreversible steps in the chain of mutations leading to cancer as comprised of substates with reversible damage attributable to DNA repair mechanisms (Kopp-Schneider, Portier, and Rippman 1991; J. A. Jacquez 1999). The last few years have seen dramatic advances in our understanding of tumor genetics, and it now is possible to sequence the genomes sampled from cancer tumors to elucidate the sequence of mutations that lead to cancer. The specific mutations vary from one tumor to another and from one patient to another, but the steps of mutation, repair, and fixation of deleterious mutations via replication events are largely the same. The substates of a model of carcinogenesis thus should be constructed to correspond to the observed tumor morphological characteristics, with flows corresponding to state transitions from mutation, repair, and replication.

Consider pancreatic cancer (Figure 3) and its corresponding compartmental model (Figure 4). A cascade of specific mutation events lead to pancreatic cancer (Alian et al. 2014), although these differ from one patient to another (Maitra and Hruban 2008). These mutations include KRAS2, p16/CDKN2A, TP53,

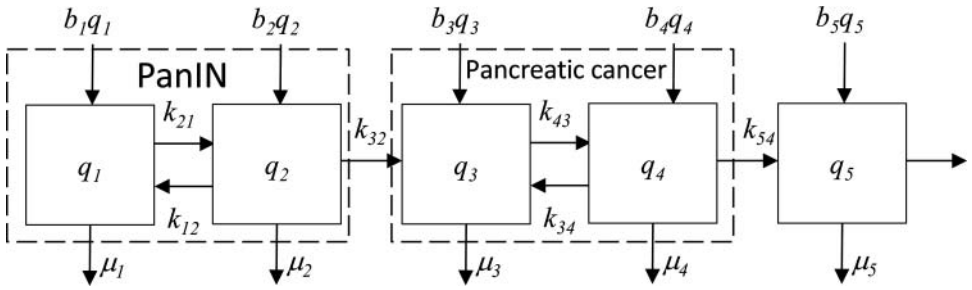

Figure 4. Model of pancreatic cancer carcinogenesis.

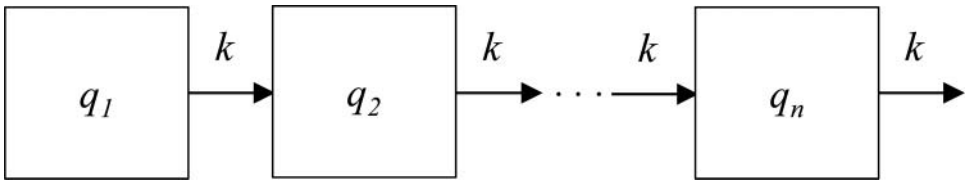

Figure 5. Outflow connected n compartment system useful for solving for the probability density function and cumulative distribution function of residence times.

SMAD4/DPC4, and other genes and result in genomic and transcriptomic alterations that lead to invasion, metastases, cell cycle deregulation, and enhanced cancer cell survival (Maitra and Hruban 2008). Precursor lesions include the mucinous cycstic neoplasm (MCN), the intraductal papillary mucinous neoplasm (IPMN), and the pancreatic intraepithelial neoplasia (PanIN). Here we consider the PanIN pathway, which is thought responsible for the majority of pancreatic cancers.

Carcinogenesis is initiated by a mutation in a normal cell that leads to accelerated cell proliferation. Waves of clonal expansion along with additional mutations progress to PanIN during time T_1 (Figure 3). This corresponds to the substates q_1 (normal cell) and q_2 (PanIN) in the model in Figure 4. One founder cell from a PanIN lesion will start the parental clone that will initiate an infiltrating carcinoma; this is indicated by the irreversible flow (k_{32}) from q_2 to q_3 in Figure 4. Here substate q_3 is the parental clone, and substate q_4 indicates subclones with metastatic capacity. The flow k_{32} to substate q_3 thus represents that replication that gives rise to the index pancreatic cancer lesion (the cells in q_3), along with the mutation events that confer metastatic capacity (resulting in the cells in q_4). The empirical residence time in substates q_3 and q_4 is T_2. The irreversible flow k_{54} indicates a proliferation and spreading of cells with metastatic capacity, with metastases (state q_5) to other organs such as the liver occurring in time T_3. The observed average times in each model state are $T_1 = 11.7$ years, $T_2 = 6.8$ years, and $T_3 = 2.7$ years (Campbell et al. 2010). These are the empirical residence times in those states describing pancreatic carcinogenesis and were estimated from tumor histology and tumor genetics. This model is consistent with recent findings regarding mechanisms of pancreatic tumorigenesis. For example, inflammation and injury are implicated as a precursor event in some pancreatic cancers, leading to acinar-to-ductal metaplasia (ADM). ADM is reversible, but an oncogenic mutation in KRAS prevents this, and the injured cells enter the pathway to PanIN. Additional mutation events then can result in pancreatic ductal adenocarcinoma (Seton-Rogers 2012), represented by the "pancreatic cancer" metacompartment in Figure 4. Details on model specification, system equations, parameter estimates, and equilibrium conditions are provided in the supplemental materials on the publisher's Web site.

Residence Times

For an outflow connected system without inflow and comprised of n compartments (Figure 5), the compartment sizes and density function of residence times, given an input of 1 unit at $t = 0$ into compartment 1, are (J. A. Jacquez 2002)

$$q_n = \frac{k^{n-1}t^{n-1}}{(n-1)!}e^{-kt} \quad (1)$$

$$\theta(\rho, t) = \frac{k^n t^{n-1}}{(n-1)!}e^{-kt}. \quad (2)$$

Here ρ specifies the proportions of particles in the n compartments such that the first compartment has size 1, and the others have size 0. This means that the initial conditions specify that all particles at time 0 are in compartment 1. These equations could be applied to solve for the density function of residence times in the compartmental model of pancreatic cancer (Figure 4) in the subsystems PanIN, pancreatic cancer, and metastatic pancreatic cancer, given certain simplifying assumptions.

See the supplemental material provided on the publisher's Web site for the estimation of residence times in the compartmental model of pancreatic cancer.

Modeling Cancer Stages: Cancer in Populations

We now present a model of pancreatic cancer stages (Figure 6). Here, the unit of observation is the cancer patient, and we observe counts of patients in early- and late-stage pancreatic cancer, before and after diagnosis (Figures 7A and 7B). Counts of people in early

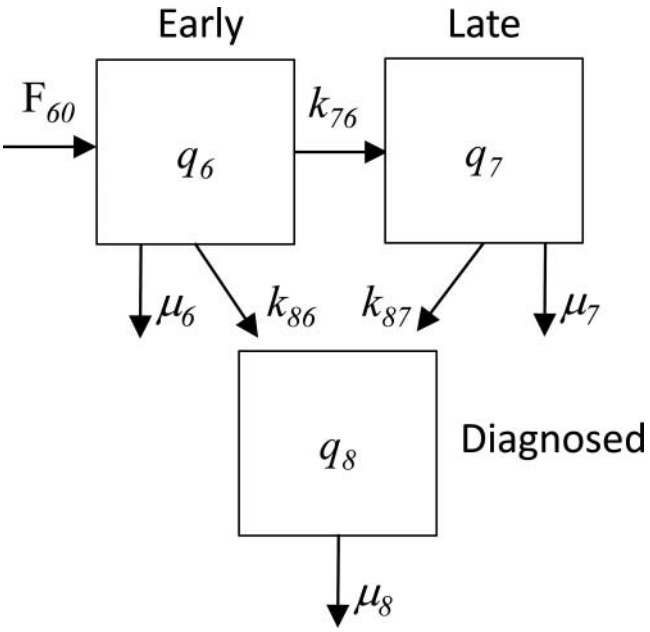

Figure 6. Stage-based model of pancreatic cancer. Here the compartment sizes are number of patients with early (q_6) and late-stage cancers (q_7) prior to diagnosis, and the number diagnosed (q_8).

and late stages prior to diagnosis are represented by q_6 and q_7. Compartment q_8 is made up of patients who have been diagnosed, in either early- or late-stage cancer. The flow of the number of healthy individuals entering early-stage cancer is F_{60}. The rate of progression from early- to late-stage cancer is k_{76}. Diagnosis events from early and late stages are given by the rates k_{86} and k_{87}. Death from the compartments is represented by μ_6, μ_7, and μ_8.

The system equations for this stage-based model of pancreatic cancer are

$$\frac{dq_6}{dt} = F_{60} - q_6(k_{76} + k_{86} + \mu_6)$$
$$\frac{dq_7}{dt} = q_6 k_{76} - q_7(k_{87} + \mu_7)$$
$$\frac{dq_8}{dt} = q_6 k_{86} + q_7 k_{87} - q_8 \mu_8. \quad (3)$$

Equilibrium occurs under the following conditions:

$$0 = \frac{dq_6}{dt} = \frac{dq_7}{dt} = \frac{dq_8}{dt}$$
$$q_6 = \frac{F_{60}}{(k_{76} + k_{86} + \mu_6)}$$

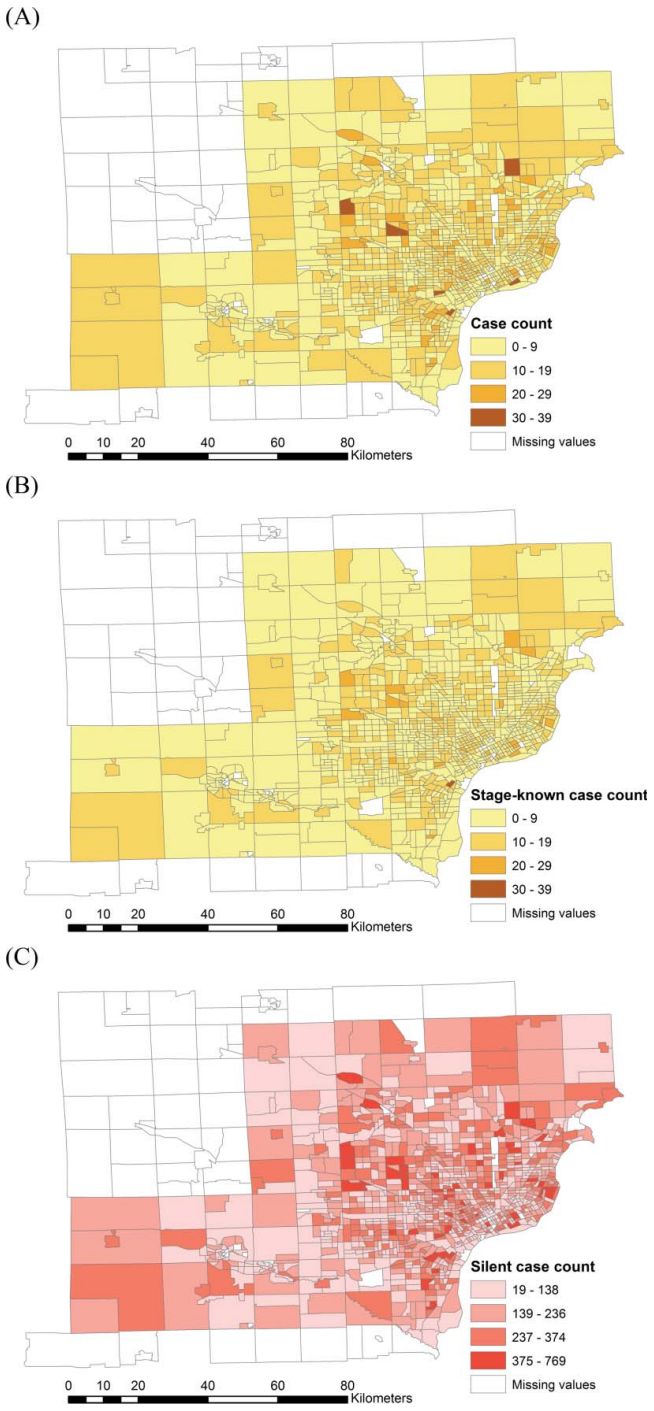

Figure 7. Choropleth maps of pancreatic cancer cases in southeast Michigan, 1985–2005: (A) incident cases; (B) stage-known cases; (C) silent (yet to be diagnosed) cases. (Color figure available online.)

$$q_7 = \frac{q_6 k_{76}}{(k_{87} + \mu_7)}$$
$$q_8 = \frac{q_6 k_{86} + q_7 k_{87}}{\mu_8}. \quad (4)$$

Table 1. Equivalence of model states and residence times between carcinogenesis- and stage-based models, using diagnostic pancreatic cancer staging according to the American Cancer Society American Joint Committee on Cancer (Edge et al. 2010)

Stage model compartment	Carcinogenesis model compartment	Description	American Joint Committee on Cancer staging	Residence time
q_6	q_3, q_4	In situ, local, not diagnosed	In situ: AJCC Tis, N0, M0 Local: AJCC IA, IB, N0, M0	T_2: 6.8 ± 3.4 years
q_7	q_5	Regional, distant, not diagnosed	Regional: AJCC IIA, IIB Distant: AJCC IV	T_3: 2.7 ± 1.2 years
q_8	—	Diagnosed pancreatic cancer	May be in situ, local, regional, or distant; in most cases pancreatic cancer is diagnosed at an advanced stage	T_4: 0.5 ± 0.25 years (2011 five-year survival rate < 6% and average life expectancy after diagnosis is 3–9 months)

Note: AJCC = American Joint Committee on Cancer; Tis = carcinoma in situ; N0 = no regional lymph node metastasis; M0 = no distant metastasis.

The number of incident early- and late-stage cancers (compartment q_8) is directly observable in most of the states in the United States from cancer registry data. The flows q_6k_{86} and q_7k_{87} are observable as the number in a defined time period of early- and late-stage diagnoses. The mortality rate $q_8\mu_8$ is directly observable as the number of diagnosed pancreatic cancer patients who die in a defined time period. The quantities $q_6\mu_6$ and $q_7\mu_7$ are the number of deaths of people with early- and late-stage undiagnosed pancreatic cancer. The estimation of parameter values and the number of yet-to-be-diagnosed cancer cases will be demonstrated later in the example of pancreatic cancer in southeast Michigan.

Table 2. Correspondence of modeled cancer stages to anatomic stages from the American Joint Committee on Cancer

Model stage	AJCC stage	Prognostic groups			Diagnosed
	Stage 0	Tis	N0	M0	N
Early (q_6)	Stage IA	T1	N0	M0	N
	Stage IB	T2	N0	M0	N
	Stage IIA	T3	N0	M0	N
Late (q_7)	Stage IIB	T1–3	N1	M0	N
	Stage III	T4	Any N	M0	N
	Stage IV	Any T	Any N	M1	N
Diagnosed (q_8)	Any stage	Any T	Any N	Any M	Y

Note: AJCC = American Joint Committee on Cancer. Primary tumor (T) coding: Tis = carcinoma in situ; T1 = tumor limited to pancreas, 2 cm or less in greatest dimension; T2 = tumor limited to pancreas, more than 2 cm in greatest dimension; T3 = tumor extends beyond the pancreas but without involvement of the celiac axis or the superior mesenteric artery; T4 = tumor involves the celiac axis or the superior mesenteric artery (unresectable primary tumor). Regional lymph nodes (N) coding: N0 = no regional lymph node metastasis; N1 = regional lymph node metastasis. Distant metastasis (M) coding: M0 = no distant metastasis; M1 = distant metastasis.

Carcinogenesis and Stage-Based Model of Pancreatic Cancer

The carcinogenesis model deals with pancreatic cancer cells in histological and genetic states as compartment members, whereas the stage-based model uses individuals and the stage of their pancreatic cancer to define compartment membership. The model of carcinogenesis informs the stage model through an equivalence of residence times and model states (Tables 1 and 2).

Application: Pancreatic Cancer in Southeast Michigan

To demonstrate the approach, we apply the stage-based model to incident pancreatic cancer cases in southeastern Michigan. We employ the four steps illustrated in Figure 2, customized to this specific application.

1. Develop the minimally sufficient biologically reasonable systems model.
2. Solve for residence times, compartment sizes, and flows.
3. Map the data to identify local populations with excess risk.
4. Interpret the results.

Background and Data

An analysis of pancreatic cancer mortality in white males in Michigan counties in two time periods from 1950 to 1970 and 1970 to 1995 found statistically significant clusters that persisted in Wayne County in

both time periods and that expanded to include adjacent Macomb County in 1970 to 1995 (G. M. Jacquez 2009). This finding was confirmed using more recent incidence and mortality data from the Surveillance Epidemiology and End Results program (SEER; Ries et al. 2007). Seventeen registry areas are included in the SEER program: Atlanta, rural Georgia, California (Bay Area, San Francisco–Oakland, San Jose–Monterey, Los Angeles, and Greater California), Connecticut, Hawaii, Iowa, Kentucky, Louisiana, New Jersey, New Mexico, Seattle-Puget Sound, Utah, and Detroit. In 2000 to 2004, Detroit had the highest age-adjusted incidence rate of pancreatic cancer for white males at 15.0 cases per 100,000 out of all of the seventeen registry areas and the second highest mortality rate at 12.9 deaths per 100,000. In contrast, the SEER-wide averages for white males in this period were 12.8 incident cases and 12.0 deaths per 100,000. Notice that the incidence is nearly equal to the deaths for the SEER-wide averages (12.8 vs. 12.0), but the incident cases in Detroit exceed the mortality rate by a larger difference (15.0 vs. 12.9). This is consistent with the observation that pancreatic cancer incidence in Detroit is increasing and that the Detroit system might not be in equilibrium. In terms of our compartmental model, it appears the flows in (F_{06}) exceed the flows out due to mortality ($q_8\mu_8$). Notably, the Detroit registry pancreatic cancer mortality for white males in 2000 to 2004 increased on average 0.9 percent per year (calculated by SEER*Stat from the National Vital Statistics System public use data file). The population covered by the Detroit registry in this period was 1,365,315 white males. The finding of excess pancreatic cancer mortality with increasing incidence was thus independently confirmed by data from SEER and found to persist from 1950 through 2004 (G. M. Jacquez 2009).

As a follow-up to this study, we obtained annual incidence data from the Michigan Cancer Surveillance Program (MCSP) for the period from 1985 to 2005. The Michigan Cancer Registry is a gold-standard registry with its completeness and accuracy certified on an annual basis, and MCSP compiles cancer records for the state. Funded in part by the National Program of Cancer Registries of the Centers for Disease Control, the MCSP is nationally certified by the North American Association of Central Cancer Registries. External audits have found a completeness percentage of 95 percent or higher on the population-based data collected by the MCSP.

Data Cleaning and Processing

The geocoding budget and numbers of observations are as follows. A total of 11,068 pancreatic cancer cases were diagnosed between 1 January 1985 and 31 December 2005. Of these, 192 addresses of place of residence at diagnosis failed to geocode, leaving 10,876 cases with known places of residence at diagnosis. Stage at diagnosis (in situ, local, regional, distant, and unknown) was recorded as unknown for 2,250 of these, leaving 8,826 cases with known place of residence and known stage at diagnosis. The head of pancreas and pancreas not otherwise specified were the most frequent primary sites, with 4,496 and 1,621 cases, respectively. Males accounted for 4,202 cases and females 4,424. By race, 6,356 cases were whites, 2,192 blacks, and the balance American Indian (8 cases), Asian (61 cases), and other or unknown groups (9 cases).

Analysis Steps

Step 1: Describe the Model

We employ the model of pancreatic cancer stages in Figures 7A and 7B and system equations in Equation 3.

Step 2: Estimate Flows, Compartment Sizes, and Residence Times

The quantities directly observable are the incident flows into compartment 8 from early- and late-stage but not diagnosed cancers. We use the data for all incident pancreatic cases, whether they were geocoded or not and whether the stage at diagnosis was known or unknown. Let o_e be the total number of cases from 1985 through 2005 observed in the early stage, o_L be the number in late stages, and o_u be the number in an unknown stage. Y is the number of years over which the observations accrued (twenty-one years). We can then estimate the flows into compartment q_8 for early- and late-stage cancers as

$$\begin{aligned}\widehat{q_6k_{86}} &= \frac{(o_e + (o_e/(o_e+o_L))o_u)}{Y} = \frac{1246.8}{21} = 59.37\\ \widehat{q_7k_{87}} &= \frac{(o_L + (o_L/(o_e+o_L))o_u)}{Y}\\ &= \frac{1246.8}{21} = 467.67.\end{aligned} \tag{5}$$

The units on these are number of cases in the given stage diagnosed per year. According to the American Cancer Society, for all stages of pancreatic cancer combined, the one-year relative survival rate is 20 percent, and the five-year rate is 4 percent. For $\mu_8 = 0.8$ deaths per diagnosed case year, and assuming the equilibrium condition in Equation 5, we estimate the size of compartment q_8 as

$$\widehat{q_8} = \frac{q_6 k_{86} + q_7 k_{87}}{\mu_8} = 658.81. \quad (6)$$

This is the average number of diagnosed and surviving (not yet deceased) pancreatic cancer cases. For the late-stage but not diagnosed cases in compartment q_7 we note that at equilibrium

$$q_7(k_{87} + \mu_7) = q_6 k_{76}. \quad (7)$$

The rate μ_7 is deaths of late-stage but not diagnosed cases that are not diagnosed after the death event and thus do not flow into compartment q_8 (they would have to be diagnosed to enter this compartment). We impose $\mu_7 = 0$, under the assumption that all of the late-stage pancreatic cancer cases are diagnosed (this assumption can be relaxed but seems reasonable because late-stage pancreatic cancers are by definition advanced and metastatic). Hence, deaths for late-stage but not yet diagnosed cases are diagnosed after they decease. This then yields

$$q_7 k_{87} = q_6 k_{76} = 467.67. \quad (8)$$

Again, the units here are number of cases per year. Because $q_7 k_{87} = q_6 k_{76}$ and $q_6 k_{86} = 59.37$,

$$\widehat{q_7} = 59.37 \frac{k_{76}}{(k_{86} + k_{87})}. \quad (9)$$

The age-adjusted annual mortality rate from all causes in Michigan in 2010 was 764.2 deaths per 100,000 (Miniño and Murphy 2012) and has decreased from 1,027.1 deaths per 100,000 in 1985 (Michigan Department of Community Health 2012). We therefore estimated the background mortality rate from 1985 to 2005 as the sum of the age-adjusted death rates for all races and sexes divided by the number of years being considered, yielding a twenty-one-year average of 924.05 deaths per 100,000. We set person-specific annual death rate $\mu_6 = 0.00924$ and using the equilibrium condition for compartment q_6 obtain

$$q_6 = \frac{F_{60}}{(k_{76} + k_{86} + \mu_6)}$$

$$F_{60} = q_6(k_{76} + k_{86} + \mu_6)$$

$$\widehat{F_{60}} = 467.67 + 59.37 + q_6\mu_6$$

$$\widehat{F_{60}} = 467.67 + 59.37 + \frac{59.37}{k_{86}}\mu_6 = 527.04 + \frac{0.5486}{k_{86}}$$

$$\widehat{q_6} = \frac{F_{60} - 527.04}{0.009241}. \quad (10)$$

Earlier we demonstrated an equivalence between residence times in early and late cancer stages (q_6 and q_7) and residence times in the carcinogenetic model of PanIN and its sequelae. Then the residence time in q_6 is T_2, and in q_7 it is T_3. It still remains to solve for the residence time in q_8, T_4. Consider a pulse of newly diagnosed cases entering q_8 either from q_6 (diagnosed in early stage) or q_7 (diagnosed in late stage). Recall that the median survival after diagnosis is six months and that the one-year survival rate is about 20 percent. Expressing time in days, we wish to fit the Erlang distribution such that CDF(182.5 days) = 0.5 and CDF (365 days) = 0.8. We solved this using the formulation for a one-compartment system with μ_8 as the exit. At a daily mortality rate of $\mu_8 = 0.0038$, we find CDF(182.5 days) = 0.5002 and CDF(365 days) = 0.7502.

Put another way, this states that for a pulse of cases diagnosed on the same day, about 50 percent will be alive after 182.5 days, and about 25 percent will be alive after one year. This indicates that our fairly simple model of compartment q_8 is reasonably complete, at least in terms of its ability to represent observed six-month and one-year survival statistics.

Now that we have estimated μ_8 we use the relationship

$$q_8 = \frac{\widehat{q_6 k_{86}} + \widehat{q_6 k_{86}}}{\mu_8} \quad (11)$$

to solve for the size of compartment 8, yielding 379.98, $\hat{q}_8 = 379.98$. This is the estimated average number of diagnosed but not deceased pancreatic cancer cases in the study area.

Earlier we solved for q_6 and q_7 using observed quantities such as incident early- and late-stage pancreatic cancer case diagnoses. It is interesting to note for q_6

that an alternative solution is to use the observed residence time in early stage, T_2, to then solve for q_6. This provides a validation of the estimate.

Define k' to be the sum of the outflow coefficients from compartment q_6, $k' = k_{76} + k_{86} + \mu_6$. Notice that we can now estimate k' using the methods developed earlier for the residence time of the Erlang distribution. Specifically, solve for k' for a one-compartment system such that the mean residence time is T_2. This yields an estimate $k' = 0.00028$, which is the per case daily rate of exit from early-stage but not yet diagnosed pancreatic cancer, attributable to background mortality, progression to advanced cancer, and diagnosis. Multiplying by q_6 and using hat notation to indicate values we can estimate from the observed data yields

$$q_6 = \widehat{q_6 k_{76}} + \widehat{q_6 k_{86}} + \widehat{q_6 \mu_6}. \tag{12}$$

We now divide through by q_6, rearrange, and have an estimator for q_6 as

$$\widehat{q_6} = \frac{\widehat{q_6 k_{76}} + \widehat{q_6 k_{86}}}{\widehat{k'} - \widehat{\mu_6}}. \tag{13}$$

Using the values obtained earlier yields (written using annual time orientation)

$$\widehat{q_6} = \frac{467.67 + 59.39}{0.1022 - 0.00924} = 5669.75. \tag{14}$$

This is the estimated number of early-stage cancers that are in the population but not yet diagnosed. We now use a similar approach to solve for the estimated number of undiagnosed advanced cancers, q_7. Recall that at equilibrium the inflows into this compartment must equal the outflows, hence $q_7 k_{87} = q_6 k_{76}$. This is estimated as the observed number of diagnosed advanced-stage cancers, and for our system $\widehat{q_7 k_{87}} = \widehat{q_6 k_{76}} = 467.67$, and $\widehat{q_6} = \frac{\widehat{q_6 k_{76}}}{\widehat{k_{87}}}$. Again, we estimate $\widehat{k_{87}}$ using the Erlang distribution of residence times. Specifically, solve for $\widehat{k_{87}}$ for a one-compartment system such that the mean residence time is T_3. This gives $\widehat{k_{87}} = 0.000703$, which is the estimated daily diagnosis rate per person with advanced-stage pancreatic cancer. Using annual values we now estimate

$$\widehat{q_7} = \frac{\widehat{q_6 k_{76}}}{\widehat{k_{87}}} = \frac{467.67}{0.257} = 1,822.6. \tag{15}$$

This is the number of individuals with undiagnosed advanced-stage pancreatic cancer.

Step 3: Map Undiagnosed Early- and Late-Stage Pancreatic Cancers; Assess Clustering of Advanced-Stage Cancers in Those Age Fifty-Five and Younger

We now estimate the numbers of undiagnosed cancers in total and for both early and late stages. We define the estimated relative risks for total undiagnosed (TRR), early-stage undiagnosed (ERR), and late-stage undiagnosed (LRR) as the proportion of cases in each of these groups (total undiagnosed, early stage undiagnosed, late stage undiagnosed) relative to the total number of diagnosed cases,

$$\widehat{TRR} = \frac{\widehat{q_6} + \widehat{q_7}}{\widehat{q_8}} = \frac{5,669.75 + 1,822.6}{379.6} = 19.72 \tag{16}$$

$$\widehat{ERR} = \frac{\widehat{q_6}}{\widehat{q_8}} = \frac{5,669.75}{379.6} = 14.92$$

$$\widehat{LRR} = \frac{\widehat{q_7}}{\widehat{q_8}} = \frac{1,822.6}{379.6} = 4.80.$$

We find that the total number of silent (yet to be diagnosed) cases is more than nineteen times the number diagnosed. Hence, for each case that is diagnosed we estimate that there are nineteen pancreatic cancer cases in the at-risk population that have yet to be diagnosed. Of these, almost fifteen are in the early stages of pancreatic cancer, and nearly five are advanced. This means that application of a screen for early-stage pancreatic cancer could dramatically reduce pancreatic cancer mortality, as such a large proportion of undiagnosed cases are in the early stages.

The choropleth maps of pancreatic cancer cases are shown in Figures 7A and 7B. The map and the frequency distribution of the estimated count of silent (yet to be diagnosed) cases are in shown in Figure 7C and Figure 8.

Step 4: Interpret Results

This analysis of pancreatic cancer in Michigan demonstrated several important findings. First, the burden

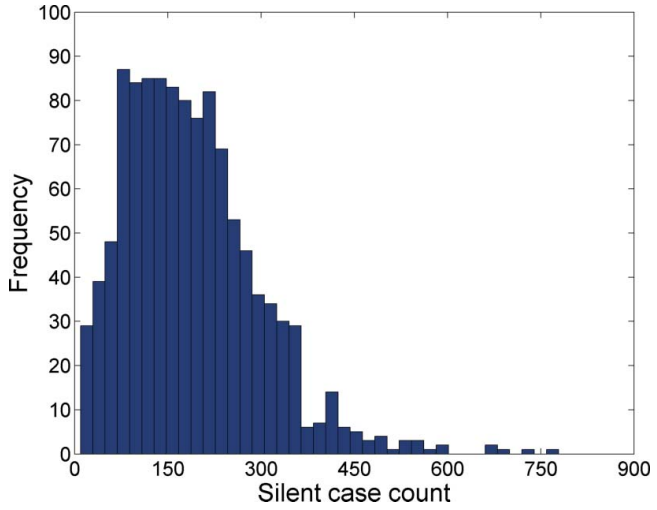

Figure 8. Frequency histogram of silent (yet to be diagnosed) pancreatic cancer cases in the greater Detroit metropolitan area. (Color figure available online.)

of undiagnosed pancreatic cancers in this population is large, approximately nineteen times the number of diagnosed pancreatic cancer cases. This indicates that a screening test for detecting early-stage pancreatic cancer, coupled with appropriate surgical and chemotherapeutic intervention, has the potential to dramatically reduce pancreatic cancer mortality in this population. Second, we estimate that there are 1,822.6 undiagnosed advanced-stage pancreatic cancer cases in this population. Some of these will be diagnosed prior to death, and others will be diagnosed postmortem. The demands on treatment resources in the last months of advanced pancreatic cancer are substantial and this estimate can be used to predict the demand for health care resources and to predict care expenses. Third, there is some evidence that pancreatic cancer risk in this population is increasing. The SEER results place pancreatic cancer incidence and mortality among the highest in all SEER registries, and the change in the annual incidence rate is about 0.9 percent per year. We found a small but statistically significant relative risk of being fifty-five or younger and late stage at diagnosis when we consider 1985 to 2005 combined. This suggests the possible action of a risk factor for pancreatic cancer that is affecting younger members of this population. Demographic factors such as differential migration cannot be excluded without further analysis, however, and in any event the relative risk is not large. Finally, the map of silent (yet to be diagnosed) pancreatic cancer cases directly supports targeting of diagnostic services, planning for upcoming in-home health care needs, and the geographic allocation of future screening programs to local populations with high demand.

Discussion

This research addresses several important topics in the modeling of space–time systems, cancer biology, and cancer surveillance. It has developed, to our knowledge, the first comprehensive modeling approach that estimates cancer latency, couples carcinogenesis and stage models, and represents and links processes at the genomic level (e.g., mutation events, cascades of genetic changes that lead to cancer), cellular level (e.g., cell replication and death, DNA repair), organ level (e.g., carcinogenesis in situ and metastases to distant organs), individual level (e.g., cancer staging in the individual, progression of individuals through cancer stages), and population level (e.g., geographic distributions of local populations in cancer stages, estimates of the predicted geographic distributions of undiagnosed cancers). Specific benefits of the approach include the following.

1. The genetic GIScience construct makes place explicit in the emerging exposome–genome+–behavome synthesis and demonstrates the vital contribution to be made by geography.
2. It is process based, capturing the known biological characteristics and mechanics of the cancer process at multiple scales (e.g., genomic to population).
3. It provides estimates of cancer latency, based on the known genetic and histologic characteristics of the cancer.
4. The latency estimates are integrated into spatiotemporal models of cancer incidence, mortality, and future cancer burden.
5. The impacts of cancer screening and diagnosis could be represented in the model by diagnosis events through which individuals progress from undiagnosed (silent) to diagnosed stages. This provides a ready mechanism for modeling improvements in pancreatic cancer screening.
6. It predicts the burden of silent cancer (yet to be diagnosed) and geographically allocates these silent cancers by cancer stages into local geographic populations. This provides the quantitative support necessary for forecasting the future cancer burden.

7. The model is readily updatable. As knowledge of cancer genomics becomes more detailed, it could be incorporated into the carcinogenesis model by updating the cascade of events that underpin the flows and stages.
8. It provides a quantitative basis for evaluating alternative treatments and for predicting treatment efficacy, provided by the equations and conditions for cancer progression, metastasis, and remission.

Several caveats apply. Assumptions implicit in compartmental models include the homogeneity assumption, which states that the particles being modeled behave in an identical fashion. This means that the pancreatic cancer cells in each compartment of the carcinogenesis model, and the cases in each compartment of the stage model, are assumed to behave in fashions identical to other particles in the compartment under consideration. This assumption is typical of all modeling approaches (because all models involve simplification and abstraction) and can be relaxed when needed by adding additional compartments to capture important aspects of heterogeneity. A second assumption of the compartmental approach is that of instantaneous and complete mixing. This assures that the kinetics (e.g., necessary for calculation of transit and residence times) of each particle can be calculated without consideration of when they entered the compartment or the order in which they entered. A final assumption is that the particles in the compartments (e.g., cells or cases) are subdividable, such that a flow of 0.3 cells is possible. This clearly is incorrect for cells and people but in practice is not a bad assumption when the number of particles in any given compartment is large.

The parameter estimates for cell replication, cell death, DNA mutation rates, repair rates, metastases initiation, cancer promotion, and so on were extracted from the literature by the first author, who is not a trained oncologist or cell biologist. Although we believe that the broad strokes are largely correct, the parameter estimates in this article are initial ones only, and the specific results might need to be revised. The overall mathematical and systems biology approach at this juncture appears sound, and it is their exposition that is the main contribution of this article (and not the initial parameter estimates).

There are several future directions for this research. First, knowledge of the exposome and its impacts on carcinogenesis could be incorporated by linking flows and coefficients related to specific exposures relevant to carcinogenetic events such as mutation, cell proliferation, replication, and other biological mechanisms through which environmental exposures affect cancer initiation and progression. For example, nonmutational mechanisms (i.e., epigenetic events that turn genes on or off through methylation) can be incorporated into the model through those model coefficients that affect tumor initiation and progression. This requires knowledge or hypotheses regarding how the epigenetic event under consideration affects carcinogenesis.

Second, the diversity of different pathways to cancer could be represented by fitting models for each pathway. For pancreatic cancers, precursor lesions include the MCN, the IPMN, and the PanIN. In this article we modeled the PanIN pathway, as it is the one responsible for the majority of pancreatic cancers. Pathway-specific models could be developed for cancers that are initiated by MCN and IPMN lesions.

Third, the carcinogenesis model provides specific conditions for cancer progression, metastasis, and remission. These could be used to predict treatment efficacy and to evaluate alternative treatments by incorporating information on how specific treatments affect those model coefficients describing cancer cell proliferation, death, and progression to distant sites. Information on how combinations of agents that differentially affect cancer cell proliferation, death, and metastatic capacity could be used in the model to evaluate novel multichemothearaputic agent treatment regimes.

Fourth, latency itself could be influenced by place-based exposure profiles. Cancer might appear earlier at higher exposures, and causative exposures might vary from one place to another. In the carcinogenesis model for the individual, this would be treated by making the mutation coefficients (presented in this article as parameters) functions based on location history. Similarly, the underlying population of undiagnosed individuals likely would have diverse exposure histories, and such heterogeneity would result in a distribution of expected times to diagnosis. The key methodology underpinning these (and other) elaborations is the ability to realistically model disease latencies, a major contribution of this article.

Finally, the technique is readily extensible to different cancers, and also to other chronic diseases.

A note on latency modeling in geographic and dynamical systems is warranted. A frequently used approach available in most dynamical system modeling software is the incorporation of specific time lags, in which the model incorporates explicit delays, in the flow from one compartment to another. Hence, one

could simply represent cancer latency by explicitly delaying (e.g., holding back) the entry of particles in the model to a destination compartment once they have exited the source compartment. This has two disadvantages. First, a priori knowledge of the time lag is required and, second, the use of explicit time lags implies the model is incomplete. When the compartmental system is properly specified, a distribution of residence times is observed that is Erlang distributed and that is representative of the empirical latency times.

A primary objective of this article has been to introduce the construct of genetic GIScience (Figure 1) and to illustrate how it could be used to inform our understanding of geographic variation in human health by incorporating knowledge of the genome+, exposome, and behavome. Geographers are largely being bypassed in the fast-moving exposome initiative. This article tries to correct this, but more needs to be done. Can geographers, for example, incorporate the social dimension more explicitly into the exposome?

The example of pancreatic cancer was used to develop process-based approaches for estimating disease latency, a key problem that must be solved for effective disease mapping and surveillance. This relied heavily on the genome+ dimension of genetic GIScience, and much work is needed to develop and exploit the exposome and behavome dimensions. The models developed can incorporate the exposome through their impact on mutation, although work is needed to make this more explicit and place based. The use of wearable sensors at the human boundary layer should prove useful in this regard. The behavome, defined as behaviors over an individual's life course that affect health, are not explicitly modeled in this article, and the quantification, representation, and modeling of the behavome is expected to be a rich future research area at the interface of human, physical, medical, and behavioral geography.

Acknowledgments

We thank Jaymie Meliker and Chantel Sloan and colleagues at the first Geolife Roundtable meeting (March 2013 in Gavle, Sweden) for thoughtful discussion and encouragement. We thank our colleagues in Australia's Cooperative Research Centers for Spatial Information (CRCSI) Health Program—Narelle Mullan, Tarun Weeamanthri, and Peter Woodgate. The first and second authors are Co-Science Directors of CRCSI Health Australia. We thank members of the Eagle Pass Yellowstone Expedition—Richard Marston, Andrew Marcus, and David Mattern for insightful criticism and comments. This avenue of research (integration of geography and population genetics) was suggested to the first author by John A. Jacquez in the 1990s.

Supplemental Material

Supplemental data for this article can be accessed on the publisher's Web site at http://dx.doi.org/10.1080/00045608.2015.1018777

References

The 1000 Genomes Project Consortium. 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature* 491 (7422): 56–65.

Alian, O. M., P. A. Philip, F. H. Sarkar, and A. S. Azmi. 2014. Systems biology approaches to pancreatic cancer detection, prevention and treatment. *Current Pharmaceutical Design* 20 (1): 73–80.

Ashton, K. 2009. That "Internet of things" thing. *RFID Journal* 22 (7): 97–114.

Ballard-Barbash, R., S. M. Siddiqi, D. A. Berrigan, S. A. Ross, L. C. Nebeling, and E. C. Dowling. 2013. Trends in research on energy balance supported by the National Cancer Institute. *American Journal of Preventive Medicine* 44 (4): 416–23.

Bentham, G. 1988. Migration and morbidity—Implications for geographical studies of disease. *Social Science & Medicine* 26 (1): 49–54.

Bian, L., Y. X. Huang, L. Mao, E. Lim, G. Lee, Y. Yang, M. Cohen, and D. Wilson. 2012. Modeling individual vulnerability to communicable diseases: A framework and design. *Annals of the Association of American Geographers* 102 (5): 1016–25.

Campbell, P. J., S. Yachida, L. J. Mudie, P. J. Stephens, E. D. Pleasance, L. A. Stebbings, L. A. Morsberger, et al. 2010. The patterns and dynamics of genomic instability in metastatic pancreatic cancer. *Nature* 467 (7319): 1109–13.

Colditz, G. A., and A. L. Frazier. 1995. Models of breast cancer show that risk is set by events of early life: Prevention efforts must shift focus. *Cancer Epidemiology Biomarkers & Prevention* 4 (5): 567–71.

Davies, K. 2010. *The $1,000 genome: The revolution in DNA sequencing and the new era of personalized medicine*. New York: Simon and Schuster.

Dearwent, S. M., R. R. Jacobs, and J. B. Halbert. 2001. Locational uncertainty in georeferencing public health datasets. *Journal of Exposure Analysis Environmental Epidemiology* 11 (4): 329–34.

Dykes, J. A., and D. M. Mountain. 2003. Seeking structure in records of spatio-temporal behaviour: Visualization issues, efforts and applications. *Computational Statistics & Data Analysis* 43 (4): 581–603.

Edge, S. B., D. R. Byrd, C. C. Compton, A. G. Fritz, F. L. Greene, and A. I. I. I. Trotti. 2010. Exocrine and endocrine pancreas. *AJCCC Cancer Staging Manual* 7:241–49.

Flintoft, L. 2014. Phenome-wide association studies go large. *Nature Reviews Genetics* 15 (2): 2.

Forer, P. 1978. A place for plastic space? *Progress in Human Geography* 2 (2): 230–67.

Griffith, D., and J. P. Paelinck. 2009. Specifying a joint space- and time-lag using a bivariate Poisson distribution. *Journal of Geographical Systems* 11 (1): 23–36.

Gulliver, J., and D. J. Briggs. 2005. Time–space modeling of journey-time exposure to traffic-related air pollution using GIS. *Environmental Research* 97 (1): 10–25.

Hägerstraand, T. 1970. What about people in regional science? *Papers in Regional Science* 24 (1): 7–24.

Hayden, E. C. 2014. Is the $1,000 genome for real? *Nature News*. http://www.nature.com/news/is-the-1-000-genome-for-real-1.14530 (last accessed 24 March 2015).

Hazin, R., K. B. Brothers, B. A. Malin, B. A. Koenig, S. C. Sanderson, M. A. Rothstein, M. S. Williams, E. W. Clayton, and I. J. Kullo. 2013. Ethical, legal, and social implications of incorporating genomic information into electronic health records. *Genetics in Medicine* 15:810–16.

Hedley, N. R., C. H. Drew, E. A. Arfin, and A. Lee. 1999. Hagerstrand revisited: Interactive space–time visualizations of complex spatial data. *Informatica-Ljubljana* 23:155–68.

Huisman, O., and P. Forer. 1998. *Computational agents and urban life spaces: A preliminary realisation of the time–geography of student lifestyles*. Paper presented at the 3rd International Conference on GeoComputation, University of Bristol, UK.

Hursting, S. 2014. Obesity, energy balance, and cancer: A mechanistic perspective. In *Advances in nutrition and cancer*, ed. V. Zappia, S. Panico, G. L. Russo, A. Budillon, and F. Della Ragione, 21–33. Berlin: Springer.

Jacquez, G. M. 2004. Current practices in the spatial analysis of cancer: Flies in the ointment. *International Journal of Health Geographics* 3 (1): 22. http://www.ij-healthgeographics.com/content/3/1/22 (last accessed 24 March 2015).

———. 2009. Cluster morphology analysis. *Spatia and Spatio-temporal Epidemiology* 1 (1): 19–29.

Jacquez, G. M., J. Barlow, R. Rommel, A. Kaufmann, M. Rienti, G. AvRuskin, and J. Rasul. 2013. Residential mobility and breast cancer in Marin County, California, USA. *International Journal of Environmental Research and Public Health* 11 (1): 271–95.

Jacquez, J. A. 1996. *Compartmental analysis in biology and medicine*. Ann Arbor, MI: Biomedware Press.

———. 1999. *Modeling with compartments*. Ann Arbor, MI: BioMedware Press.

———. 2002. Density functions of residence times for deterministic and stochastic compartmental systems. *Mathematical Biosciences* 180:127–39.

Jacquez, J. A., and C. Simon, 2002. Qualitative theory of compartmental systems with lags. *Mathematical Biosciences* 180 (1): 329–62.

Kho, A. N., L. V. Rasmussen, J. J. Connolly, P. L. Peissig, J. Starren, H. Hakonarson, and M. G. Hayes. 2013. Practical challenges in integrating genomic data into the electronic health record. *Genetics in Medicine* 15:772–78.

Kopp-Schneider, A., C. J. Portier, and F. Rippman. 1991. The application of a multistage model that incorporates DNA damage and repair to the analysis of initiation/promotion experiments. *Mathematical Biosciences* 105:139–66.

Kwan, M. P. 2000. Interactive geovisualization of activity-travel patterns using three-dimensional geographical information systems: A methodological exploration with a large data set. *Transportation Research Part C: Emerging Technologies* 8 (1–6): 185–203.

———. 2009. From place-based to people-based exposure measures. *Social Science & Medicine* 69 (9): 1311–13.

———. 2012. The uncertain geographic context problem. *Annals of the Association of American Geographers* 102 (5): 958–68.

Lam, N. S. N. 2012. Geospatial methods for reducing uncertainties in environmental health risk assessment: Challenges and opportunities. *Annals of the Association of American Geographers* 102 (5): 942–50.

Lioy, P. J., and S. M. Rappaport. 2011. Exposure science and the exposome: An opportunity for coherence in the environmental health sciences. *Environmental Health Perspectives* 119 (11): A466–A467.

Lowe, S. A., and G. OLaighin. 2014. Monitoring human health behaviour in one's living environment: A technological review. *Medical Engineering & Physics* 36 (2): 147–68.

Maitra, A., and R. H. Hruban. 2008. Pancreatic cancer. *Annual Review of Pathology: Mechanisms of Disease* 3 (1): 157–88.

Marmot, M. 2000. Social determinants of health: From observation to policy. *Medical Journal of Australia* 172 (8): 379–82.

Mayne, D., G. Morgan, A. Willmore, N. Rose, B. Jalaludin, H. Bambrick, and A. Bauman. 2013. An objective index of walkability for research and planning in the Sydney Metropolitan Region of New South Wales, Australia: An ecological study. *International Journal of Health Geographics* 12 (1): 61. http://www.ij-healthgeographics.com/content/12/1/61 (last accessed 24 March 2015).

Meade, M. S. 1977. Medical geography as human ecology: The dimension of population movement. *Geographical Review* 67:379–93.

Mefford, H. C. 2012. Diagnostic exome sequencing—Are we there yet? *New England Journal of Medicine* 367 (20): 1951–53.

Meissner, A. 2012. What can epigenomics do for you? *Genome Biology* 13 (10): 420.

Michigan Department of Community Health. 2012. Age-adjusted death rates by race and sex Michigan and United States residents, 1980–2012. http://www.mdch.state.mi.us/pha/osr/deaths/dxrates.asp (last accessed 21 May 2014).

Miller, G. W., and D. P. Jones. 2014. The nature of nurture: Refining the definition of the exposome. *Toxicological Sciences* 137 (1): 1–2.

Miller, H. J. 1999. Measuring space-time accessibility benefits within transportation networks: Basic theory and

computational procedures. *Geographical Analysis* 31 (2): 187–212.

Miniño, A. M., and S. L. Murphy. 2012. Death in the United States, 2010. In *NCHS data brief*. Hyattsville, MD: National Center for Health Statistics.

Nikiforov, Y., and D. R. Gnepp. 1994. Pediatric thyroid cancer after the Chernobyl disaster: Pathomorphologic study of 84 cases (1991–1992) from the republic of Belarus. *Cancer* 74 (2): 748–66.

Orešič, M. 2014. Systems biology in human health and disease. In *A systems biology approach to study metabolic syndrome*, ed. M. Orešič; and A. Vidal-Puig, 17–23. Berlin: Springer.

Palumbo, F., P. Barsocchi, C. Gallicchio, S. Chessa, and A. Micheli. 2013. Multisensor data fusion for activity recognition based on reservoir computing. In *Evaluating AAL systems through competitive benchmarking*, ed. J. Botía, J. Álvarez-García, K. Fujinami, P. Barsocchi, and T. Riedel, 24–35. Berlin: Springer.

Picheral, H. 1982. Géographie médicale, géographie des maladies, géographie de lu santé [Medical geography, the geography of diseases, the geography of health]. *Espace géographique* 11 (3): 161–75.

Pickle, L. W., L. A. Waller, and A. B. Lawson. 2005. Current practices in cancer spatial data analysis: A call for guidance. *International Journal of Health Geographics* 4 (1): 3.

Pybus, M., G. M. Dall'Olio, P. Luisi, M. Uzkudun, A. Carreño-Torres, P. Pavlidis, H. Laayouni, J. Bertranpetit, and J. Engelken. 2014. 1000 Genomes Selection Browser 1.0: A genome browser dedicated to signatures of natural selection in modern humans. *Nucleic Acids Research* 42 (D1): D903–D909.

Rappaport, S. M. 2011. Implications of the exposome for exposure science. *Journal of Exposure Science and Environmental Epidemiology* 21 (1): 5–9.

Richardson, D. B., N. D. Volkow, M.-P. Kwan, R. M. Kaplan, M. F. Goodchild, and R. T. Croyle. 2013. Spatial turn in health research. *Science* 339 (6126): 1390–92.

Ries, L. A. G., D. Melbert, M. Krapcho, A. Mariotto, B. A. Miller, E. J. Feuer, L. Clegg, et al. 2007. SEER cancer statistics review, 1975–2004. http://seer.cancer.gov/archive/csr/1975_2004/(lastaccessed21May2014).

Rothman, K. J. 1981. Induction and latent periods. *American Journal of Epidemiology* 114 (2): 253–59.

Sabel, C. E., P. J. Boyle, M. Loytonen, A. C. Gatrell, M. Jokelainen, R. Flowerdew, and P. Maasilta. 2003. Spatial clustering of amyotrophic lateral sclerosis in Finland at place of birth and place of death. *American Journal of Epidemiology* 157 (10): 898–905.

Sabel, C. E., A. C. Gatrell, M. Loytonen, P. Maasilta, and M. Jokelainen. 2000. Modelling exposure opportunities: Estimating relative risk for motor neurone disease in Finland. *Social Science & Medicine* 50 (7–8): 1121–37.

Schoech, D., J. F. Boyas, B. M. Black, and N. Elias-Lambert. 2013. Gamification for behavior change: Lessons from developing a social, multiuser, web-tablet based prevention game for youths. *Journal of Technology in Human Services* 31 (3): 197–217.

Seton-Rogers, S. 2012. Tumorigenesis: Pushing pancreatic cancer to take off. *Nature Reviews Cancer* 12 (11): 739.

Sinha, G., and D. Mark. 2005. Measuring similarity between geospatial lifelines in studies of environmental health. *Journal of Geographical Systems* 7 (1): 115–36.

Swan, M. 2012. Sensor mania! The Internet of things, wearable computing, objective metrics, and the quantified self 2.0. *Journal of Sensor and Actuator Networks* 1: 217–53.

Tarczy-Hornoch, P., L. Amendola, S. J. Aronson, L. Garraway, S. Gray, R. W. Grundmeier, L. A. Hindorff, et al. 2013. A survey of informatics approaches to whole-exome and whole-genome clinical reporting in the electronic health record. *Genetics in Medicine* 15:824–32.

Tunstall, H. V. Z., M. Shaw, and D. Dorling. 2004. Places and health. *Journal of Epidemiology and Community Health* 58 (1): 6–10.

van Tongeren, M., and J. W. Cherrie. 2012. An integrated approach to the exposome. *Environmental Health Perspectives* 120 (3): A103–A104.

Weinhold, B. 2012. More chemicals show epigenetic effects across generations. *Environmental Health Perspectives* 120 (6): a228.

Wheeler, D. C., M. H. Ward, and L. A. Waller. 2012. Spatial–temporal analysis of cancer risk in epidemiologic studies with residential histories. *Annals of the Association of American Geographers* 102 (5): 1049–57.

Whitson, J. R. 2013. Gaming the quantified self. *Surveillance & Society* 11 (1–2): 163–76.

Wild, C. P. 2005. Complementing the genome with an "exposome": The outstanding challenge of environmental exposure measurement in molecular epidemiology. *Cancer Epidemiology Biomarkers & Prevention* 14 (8): 1847–50.

———. 2012. The exposome: From concept to utility. *International Journal of Epidemiology* 41 (1): 24–32.

Yachida, S., S. Jones, I. Bozic, T. Antal, R. Leary, B. Fu, M. Kamiyama, et al. 2010. Distant metastasis occurs late during the genetic evolution of pancreatic cancer. *Nature* 467 (7319): 1114–17.

Yu, J.-H., S. M. Jamal, H. K. Tabor, and M. J. Bamshad. 2013. Self-guided management of exome and whole-genome sequencing results: Changing the results return model. *Genetics in Medicine* 15:684–90.

Zentner, G., and S. Henikoff. 2012. Surveying the epigenomic landscape, one base at a time. *Genome Biology* 13 (10): 250.

Correspondence: Department of Geography, University at Buffalo–State University of New York, Buffalo, NY 14261, e-mail: gjacquez@buffalo.edu (Jacquez); chenshi@buffalo.edu (Shi); School of Geographical Sciences, University of Bristol, Clifton, Bristol BS8 1SS, UK, e-mail: c.sabel@bristol.ac.uk (Sabel).