

The spread of hepatitis C virus genotype 1a in North America: a retrospective phylogenetic study



Jeffrey B Joy, Rosemary M McCloskey, Thuy Nguyen, Richard H Liang, Yury Khudyakov, Andrea Olmstead, Mel Krajden, John W Ward, P Richard Harrigan, Julio S G Montaner, Art F Y Poon

Summary

Background The timing of the initial spread of hepatitis C virus genotype 1a in North America is controversial. In particular, how and when hepatitis C virus reached extraordinary prevalence in specific demographic groups remains unclear. We quantified, using all available hepatitis C virus sequence data and phylodynamic methods, the timing of the spread of hepatitis C virus genotype 1a in North America.

Methods We screened 45 316 publicly available sequences of hepatitis C virus genotype 1a for location and genotype, and then did phylogenetic analyses of available North American sequences from five hepatitis C virus genes (*E1*, *E2*, *NS2*, *NS4B*, *NS5B*), with an emphasis on including as many sequences with early collection dates as possible. We inferred the historical population dynamics of this epidemic for all five gene regions using Bayesian skyline plots.

Findings Most of the spread of genotype 1a in North America occurred before 1965, and the hepatitis C virus epidemic has undergone relatively little expansion since then. The effective population size of the North American epidemic stabilised around 1960. These results were robust across all five gene regions analysed, although analyses of each gene separately show substantial variation in estimates of the timing of the early exponential growth, ranging roughly from 1940 for *NS2*, to 1965 for *NS4B*.

Interpretation The expansion of genotype 1a before 1965 suggests that nosocomial or iatrogenic factors rather than past sporadic behavioural risk (ie, experimentation with injection drug use, unsafe tattooing, high risk sex, travel to high endemic areas) were key contributors to the hepatitis C virus epidemic in North America. Our results might reduce stigmatisation around screening and diagnosis, potentially increasing rates of screening and treatment for hepatitis C virus.

Funding The Canadian Institutes of Health Research, Michael Smith Foundation for Health Research, and BC Centre for Excellence in HIV/AIDS.

Introduction

Hepatitis C virus is a global threat to public health with an estimated 185 million people infected worldwide, including 4·6 million infections of predominantly genotype 1a in North America.¹ The North American epidemic is composed of 3·5 million infections in the USA, 300 000 infections in Canada, and 900 000 in Mexico.¹ Infected individuals are at high risk of liver cirrhosis, hepatocellular carcinoma, and liver failure, leading to early mortality and high cost to the health-care system. About 75% of adults infected with hepatitis C virus in North America were born between 1945 and 1964 (so-called baby boomers).² The mechanism by which hepatitis C virus reached such high prevalence in this cohort is unclear. Previous studies have implicated the use of infected blood products before the advent of rigorous screening of the blood supply for infectious agents in the early 1990s, or injection drug use, which peaked in North America at the end of the 1960s.^{3,4} However, despite mixed evidence in the published work, the dominant view is that the baby boomer epidemic in North America is largely attributable to past sporadic risky behaviours (ie, experimentation with injection drug use, unsafe tattooing, high risk sex, travel to high prevalence areas)^{2,5,6} and transfusion with infected blood.

Many RNA viruses, including hepatitis C virus, evolve rapidly and therefore it is possible to approximate the recent transmission history of an epidemic from the genetic divergence of infections sampled from the population.⁷ This genetic diversity is structured by the shared ancestries of the infections, which can be reconstructed from these sequences as a tree-based model known as a phylogeny. Since most of the hepatitis C virus genome evolves at roughly 0·001 substitutions per site per year,^{8,9} it is possible to use the collection dates of different samples to rescale the phylogeny such that the branching points in the tree approximate the dates of transmission events in the epidemic.¹⁰ Thus, the shape of the phylogeny contains valuable information about the epidemiological dynamics of the virus population,^{11,12} which can be extracted by fitting epidemic or population genetic models to these data. Such techniques have provided key insights into epidemics including HIV, influenza A virus, dengue virus, and Ebola virus infection.^{4,7,13} Many of these studies focused on the effective number of infections, which is a quantity expected to be proportional to prevalence during the exponential phase of an epidemic, but which should not in general be mistaken as equivalent to prevalence.¹⁴

Lancet Infect Dis 2016

Published Online
March 30, 2016
[http://dx.doi.org/10.1016/S1473-3099\(16\)30002-0](http://dx.doi.org/10.1016/S1473-3099(16)30002-0)

See Online/Comment
[http://dx.doi.org/10.1016/S1473-3099\(16\)30002-0](http://dx.doi.org/10.1016/S1473-3099(16)30002-0)

BC Centre for Excellence in HIV/AIDS, Vancouver, BC, Canada (J B Joy PhD, R M McCloskey BSc, T Nguyen BSc, R H Liang PhD, P R Harrigan PhD, J S G Montaner MD, A F Y Poon PhD); Department of Medicine (J B Joy, P R Harrigan, J S G Montaner, A F Y Poon), Department of Pathology and Laboratory Medicine (M Krajden MD), University of British Columbia, Vancouver, BC, Canada; Division of Viral Hepatitis, Centers for Disease Control and Prevention, Atlanta, GA, USA (Y Khudyakov PhD, J W Ward MD); and BC Centre for Disease Control, Vancouver, BC, Canada (A Olmstead PhD, M Krajden)

Correspondence to:
Dr Jeffrey B Joy, BC Centre for Excellence in HIV/AIDS, Vancouver, BC V6Z 1Y6, Canada
jjoy@cfenet.ubc.ca

Research in context**Evidence before this study**

We searched Scopus, PubMed, and Google Scholar for articles published up to Dec 2, 2015, in any language, that addressed the spread of hepatitis C virus in North America and globally. Most adults infected with hepatitis C virus in North America, and some other high-income regions, are in the cohort of individuals born between 1945 and 1964 (so-called baby boomers). How the virus reached such high prevalence in this cohort remains unclear. Previous studies implicated both the use of infected blood products before screening of the blood supply in the early 1990s and injection drug use, which peaked in North America at the end of the 1960s. However, despite mixed evidence, the epidemic in North America is thought to be largely attributable to past sporadic risky behaviours (ie, experimentation with injection drug use, unsafe tattooing, high risk sex, travel to high prevalence areas), along with limited transmission through transfusion with infected blood products.

Added value of this study

Our results, based on careful phylogenetic analyses within a large sample of North American hepatitis C virus sequences,

revise previous estimates of the exponential phase of growth of the North American genotype 1a epidemic back by more than 15 years to roughly 1950. These results dispute the idea that the epidemic among baby boomers and other demographic groups in North America is primarily due to injection drug use, unsafe tattooing, high risk sex, and travel to high endemic areas during youth. It instead suggests that the epidemic is linked to both the increase in medical procedures that occurred during and following World War 2 and the use of reusable glass and metal syringes in the 1950s and 1960s and subsequently risky behaviours.

Implications of all the available evidence

The spread of hepatitis C virus genotype 1a in North America and dominance in baby boomers was a result of early expansion because of both iatrogenic and risky behaviours, followed by transmission due to intravenous drug use in later years as medical technology and blood supply screening prevented iatrogenic transmission.

In this retrospective longitudinal study, we used a distribution of phylogenetic trees of all published sequences covering five hepatitis C virus genes to infer the past trajectories in the effective number of infections in the North American genotype 1a epidemic. We focused on the genotype 1a epidemic because it is overwhelmingly the dominant hepatitis C virus genotype circulating in North America.¹ We assessed the progression of the epidemic over time to test whether the timing of the spread of hepatitis C virus in North America matches the previously hypothesised periods of high incidence. If the dynamics of the hepatitis C virus epidemic in North America have primarily been driven by injection drug use or the use of infected blood products, then we would expect the effective number of infections to be highest later in the progression of the epidemic—namely, at the end of the 1960s and early 1970s.

Methods**Data collection**

We retrieved all available records of hepatitis C virus sequences from the National Center for Biotechnology Information Genbank nucleotide database (appendix p 12). The search results, consisting of 160 556 records, were reduced to 45 316 sequences by excluding any sequence lacking a source modifier field (attribute information associated with the sequence) or a collection date.

Specific gene regions were extracted from these data by pairwise alignment of each sequence against the respective sequence intervals from the H77 reference genome and extracting the overlapping region. Each gene-specific dataset was manually inspected and clipped to extract the

broadest interval with the greatest overlap across sequences. We selected genomic regions for analyses on the basis of previous whole genome studies of hepatitis C virus genotype 1, which showed that *E2*, *NS2*, and *NS5B* were the most informative gene regions for phylodynamic studies.⁴ We also analysed the *E1* and *NS4B* regions.⁴

Data curation

To screen for genotype 1a sequences, genotype and gene-specific reference sequences were collected from the Los Alamos National Laboratory hepatitis C virus sequence repository. We aligned sequences in each dataset using MAFFT (version 7.154b) and manually inspected them using HYPHY (version 2.22). Phylogenetic trees for each complete dataset were inferred in an approximate maximum likelihood modelling framework as implemented in FastTree2, then visualised and inspected with FigTree to isolate the clade comprising all sequences annotated as genotype 1a. This procedure identified 13 sequences whose GenBank records were incorrectly annotated with a different genotype. Subsequently, the resulting hepatitis C virus gene-specific datasets consisted of 917 sequences in total and covered 4499 base pairs (*E1*: 252 sequences, 576 bp; *E2*: 196 sequences, 1089 bp; *NS2*: 182 sequences, 651 bp; *NS4B*: 139 sequences, 783 bp; *NS5B*: 148 sequences, 1400 bp). The appendix shows accession numbers for all public domain sequences used (pp 5–10). Included sequences were sampled between 1977 and 2011 (appendix p 13). Approximate maximum likelihood phylogenetic trees for each gene region are shown in the appendix (p 14).

For more on FigTree see
[http://tree.bio.ed.ac.uk/
software/figtree/](http://tree.bio.ed.ac.uk/software/figtree/)

See Online for appendix

Effective number of infections

We reconstructed the dynamics of the epidemic using Bayesian skyline plots, a non-parametric smoothing method for approximating past population dynamics.¹⁵ The y axis in a Bayesian skyline plot represents N_e multiplied by τ , where N_e is the effective population size and τ is the generation time. The y axis can be interpreted as the number of infected individuals who go on to infect additional individuals (effective number of infections¹⁶), and the x axis represents time. A minimum of four replicate Markov chain Monte Carlo samples were run in BEAST (version 1.8.0) for each gene, for at least 5×10^8 generations per run. To determine if choice of priors on both the model by which sequences accrue changes (substitution models) or the model by which the expected number of changes in a sequence relates to time (clock models) biased the results, we did two runs for each gene with four different substitution models (General Time Reversible and Tamura-Nei 93 with and without γ distributed rate variation) and three different clock models (strict, random local, and relaxed). To determine the best fitting model, we compared models in the programme Tracer (version 1.5) using Bayes factors.¹⁰ The optimum run conditions were a General Time Reversible substitution model with among-site rate heterogeneity distributed according to a γ distribution under a relaxed molecular clock. The convergence of each run was assessed through evaluation of parameter traces and effective sample size greater than 200 via Tracer (version 1.5).

Role of the funding source

The funders of this study had no role in the study design, data collection, data analysis, data interpretation, or writing of the report. The corresponding author had full access to all the data in the study and had final responsibility for the decision to submit for publication.

Results

The figure shows Bayesian skyline plots reconstructing the spread of the genotype 1a epidemic of hepatitis C virus in North America for all analysed genomic regions; the appendix (p 16) shows each individual gene region. Taken together, the analyses of all gene regions suggest that the greatest expansion of the epidemic in North America occurred between 1940 and 1965. The exponential growth phase of the epidemic had subsided by 1965, suggesting a plateau in the rate of spread between 1965 and 1989, then dropped during the early 1990s. There has been a slight increase in the effective number of hepatitis C virus infections since the early 2000s (figure). Analyses of each gene separately show substantial variation in estimates of the timing of the early exponential growth, ranging roughly from 1940 for NS2, to 1965 for NS4B (figure). We had the greatest confidence in the peak spread across all genes between 1948 and 1963. In sum, our phylogenetic analyses

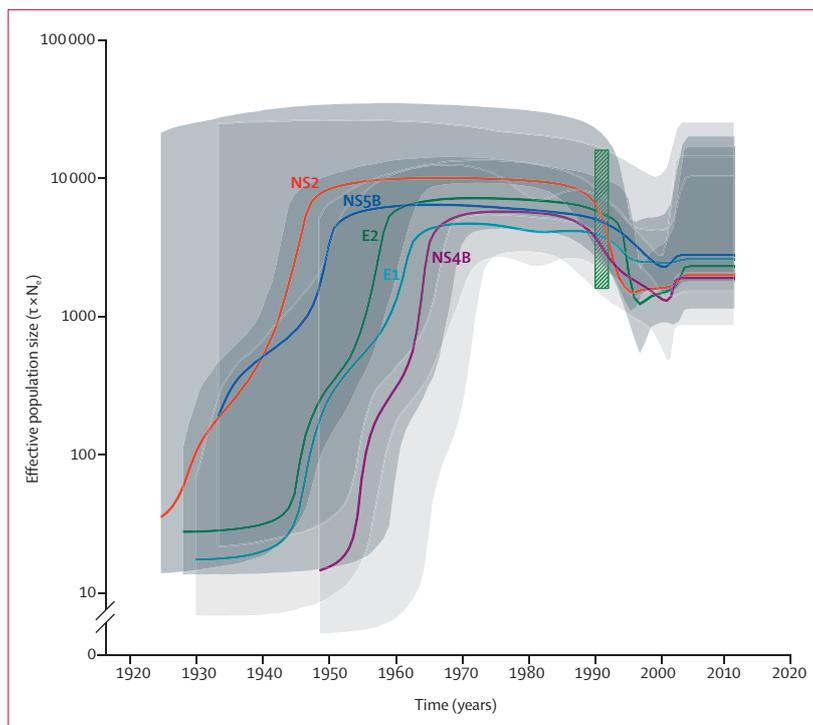


Figure: Median estimated hepatitis C virus effective number of infections in North America

Coloured lines represent the median estimated effective number of infections (effective number of infections \times generation time; $N_e \tau$) through time estimated under a relaxed-clock model. The 95% highest probability density for each gene region is shaded in grey. Areas of darker grey represent areas with the most highest probability density overlap across gene regions. The vertical hatched green box corresponds to the advent of both rigorous screening of the blood supply for infectious agents in 1992,¹⁷ and the widespread availability of needle exchange programmes in North America.¹⁸

strongly suggest that the hepatitis C virus genotype 1a epidemic in North America had already attained the height of its distribution by 1960.

For Tracer see <http://tree.bio.ed.ac.uk>

Discussion

Our results suggest that the hepatitis C virus genotype 1a epidemic underwent rapid expansion between 1940 and 1960. There followed a period of relatively little growth in the epidemic, a subsequent decline during the early 1990s, and a small increase in the late 1990s and early 2000s. These analyses suggest the period of greatest increase in North America was substantially earlier than previously suggested.⁴

Potential causes of a substantial early expansion (1940–60) of the epidemic are diverse, coinciding with the increase in medical procedures during and immediately following World War 2,^{4,19,20} when injection and blood transfusion technologies were still in their infancy,^{20,21} and with the expansion of recreational injection drug use in North America and associated needle sharing between 1920 and the late 1960s.^{3,5} Many medical procedures have been linked to the spread of hepatitis C virus including use of contaminated multidose vials,^{22–24} finger stick devices,²⁵ and surgical procedures.²⁶ The spread of hepatitis C virus in North America through iatrogenic

sources between 1940 and 1960 is consistent with evidence of iatrogenic spread in other high-income regions over the same period—eg, in Russia,²⁷ France,²⁸ and Japan.^{20,27} Injection drug use became more common in North America in the late 1960s, suggesting that if this was the main driver of the epidemic, the period of peak expansion of the hepatitis C virus epidemic would have coincided with the peak of injection drug use rather than much earlier as shown by our analyses.

The plateau in the spread of hepatitis C virus between 1960 and 1990 is consistent with the hypothesis that changes in injection technology were a driving factor. Before 1950, injection technology was characterised by machine-made glass and metal syringes, which were typically sterilised manually and reused because of their high cost.²¹ Between 1950 and 1960, such syringes were phased out and replaced by disposable plastic single-use syringes.²¹ Since 1960, reuse of medical syringes has been greatly reduced in North America.²¹ The spread of hepatitis C virus during this plateau period could have been maintained by a combination of injection drug use^{3,4} and transfusion of infected blood products before the introduction of rigorous screening of blood supplies in 1992.¹⁷ Before 1992, the incorporation of plasma infected with hepatitis C virus into the blood supply often resulted in nosocomial transmission during medical procedures involving the transfusion of blood or organ transplantation.¹⁷ However, transmission of hepatitis C virus via infected blood products combined with transmission through injection drug use probably occurred at a rate sufficient to maintain the effective population size of hepatitis C virus.^{3,5,6} This hypothesis is further supported by the decline in spread in the early 1990s coincident with the advent of rigorous screening of blood products¹⁷ and with the widespread penetration of needle exchange programmes into communities of injection drug users.¹⁸ The slight upturn in effective population size just after the year 2000 is consistent with epidemiological evidence of both increases in hepatitis C virus infection among young injection drug users who reside in non-urban areas,²⁹ and a significant increase in infections among men who have sex with men.³⁰

Because of the paucity of samples with early collection dates, previous studies of genotype 1a population dynamics used sequences collected mainly between 2000 and 2010. For example, a landmark study⁴ used one sequence sampled before 1980 (H77), six before 1995, 44 between 2000 and 2004, and 35 between 2005 and 2007 (appendix p 15). Thanks to improvements in sequencing technology and the technical ability to work with archival samples, more data with earlier collection dates are now available, enabling a more even distribution of sampling times (appendix pp 11, 13, and 15).

A similar analysis by Magiorkinis and colleagues,⁴ which focused on the *NS5B* gene, estimated the main expansion of the global genotype 1a epidemic to be

between 1960 and 1980, peaking in 1975. Our reanalysis of the data of Magiorkinis and colleagues, using similar methods but restricted to samples collected in the USA, was consistent with their findings (data not shown). The discordance between the timing of the spread of genotype 1a in North America in our study and that by Magiorkinis and colleagues most likely reflects the greater number of early sequences included in our analyses (appendix pp 11, 13, and 15) and the fact that our study was restricted to North American sequences whereas Magiorkinis and colleagues included sequences from other regions. The rate at which changes in nucleotide sequences (substitutions) accrue differs between lineages and over time. Changes over time might occur because of fluctuations in population size, host-specific environment effects, the intensity of natural selection, changes in protein function, or changes in generation times. Thus, if most samples are from later periods, this can skew estimates in the direction of when most samples were obtained. A more even distribution of sampling times will provide more robust estimates that are less affected by such variations in the rate at which molecular changes accumulate in genetic sequences.

Based on our results, the oldest members of the demographic cohort with the highest burden of hepatitis C virus (the baby boomers) were roughly 5 years of age around the peak of the spread of genotype 1a in North America in 1950. Thus, it is unlikely that past sporadic risky behaviour (experimentation with injection drug use, unsafe tattooing, high risk sex, travel to endemic areas) was the dominant route of transmission in this group. However, infection with hepatitis C virus in this cohort certainly occurred through injection drug use thereafter, and remains a dominant source of new infections.^{5,6}

Limitations of this analysis include the fact that phylogenetic methods of dating ancestors from genetic sequences are sensitive to invalid model assumptions¹³ making it difficult to confidently pinpoint a difference of 10–20 years in the past. The inclusion of more older sequences obtained from archived samples would be of great value in refining estimates of the timing of the spread while improving confidence intervals around the estimates. It is not unusual for molecular clock estimates to have broad confidence intervals when reconstructing events deep in the tree; furthermore, our gene-specific datasets did not overlap completely. Thus, we were integrating sequences derived from different genomes that have been subjected to different environments. We should not therefore expect the confidence limits on these different gene-specific datasets to be highly concordant in the remote past.

In summary, our data indicate that the rapid and large-scale expansion of hepatitis C virus transmission in North America was coincident with increases in medical procedures that began after World War 2 (nosocomial or iatrogenic causes), and not only the rise

in injection drug use, which peaked much later in North America, in the late 1960s.³ Thus, the prevailing view that the North American epidemic is predominately attributable to past sporadic risky behaviours is not supported by our data. Belief that risky behaviours were the dominant route of infection in North America could bias providers and patients against screening. Our results could help to reduce stigma related to screening and diagnosis of hepatitis C virus, potentially increasing the numbers of providers offering screening, patients accepting testing, and if positive, presenting for care and treatment. Because of vast improvements in medical technology, particularly in syringe manufacture and guidelines for their use, nosocomial factors that contributed to the early spread of hepatitis C virus in North America⁴ no longer have a major role in the epidemic.

Contributors

JB and AFYP designed the study. JBJ, TN, RMM, and AFYP acquired data. JBJ, RHL, and AFYP analysed and interpreted the data. JBJ and AFYP drafted the report. All authors contributed to the critical revision of the report and approved the final version.

Declaration of interests

We declare no competing interests.

References

- Messina JP, Humphreys I, Flaxman A, et al. Global distribution and prevalence of hepatitis C virus genotypes. *Hepatology* 2015; **61**: 77–87.
- Cohen J. Calling all baby boomers: get your hepatitis C test. *Science* 2012; **337**: 903.
- Courtwright DT, Courtwright DT. Dark paradise: a history of opiate addiction in America. Harvard University Press, 2009.
- Magiorkinis G, Magiorkinis E, Paraskevis D, et al. The global spread of hepatitis C virus 1a and 1b: a phylogenetic and phylogeographic analysis. *PLoS Med* 2009; **6**: e1000198.
- Williams IT, Bell BP, Kuhnert W, Alter MJ. Incidence and transmission patterns of acute hepatitis C in the United States, 1982–2006. *Arch Intern Med* 2011; **171**: 242–48.
- Alter MJ. Epidemiology of hepatitis C. *Hepatology* 1997; **26**: 62S–65S.
- Poon AFY, Joy JB, Woods CK, et al. The impact of clinical, demographic and risk factors on rates of HIV transmission: a population-based phylogenetic analysis in British Columbia, Canada. *J Infect Dis* 2014; **211**: 926–35.
- Gray RR, Parker J, Lemey P, Salemi M, Katzourakis A, Pybus OG. The mode and tempo of hepatitis C virus evolution within and among hosts. *BMC Evol Biol* 2011; **11**: 131.
- Kimura M. Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. *Nature* 1977; **267**: 275–76.
- Rodrigo AG, Felsenstein J. Coalescent approaches to HIV population genetics. The evolution of HIV. Baltimore: Johns Hopkins University Press, 1999: 233–72.
- Stadler T, Bonhoeffer S. Uncovering epidemiological dynamics in heterogeneous host populations using phylogenetic methods. *Philos Trans R Soc Lond B Biol Sci* 2013; **368**: 20120198.
- Holmes EC, Nee S, Rambaut A, Garnett GP, Harvey PH. Revealing the history of infectious disease epidemics through phylogenetic trees. *Philos Trans R Soc Lond B Biol Sci* 1995; **349**: 33–40.
- Worobey M, Rambaut A, Han G-Z. A synchronized global sweep of the internal genes of modern avian influenza virus. *Nature* 2014; **508**: 254–57.
- Frost SDW, Volz EM. Viral phylodynamics and the search for an 'effective number of infections'. *Philos Trans R Soc Lond B Biol Sci* 2010; **365**: 1879–90.
- Drummond AJ, Rambaut A, Shapiro B, Pybus OG. Bayesian coalescent inference of past population dynamics from molecular sequences. *Mol Biol Evol* 2005; **22**: 1185–92.
- Frost SDW, Volz EM. Modelling tree shape and structure in viral phylodynamics. *Philos Trans R Soc Lond B Biol Sci* 2013; **368**: 1614.
- Rich JD, Taylor LE. The beginning of a new era in understanding hepatitis C virus prevention. *J Infect Dis* 2010; **202**: 981–83.
- Des Jarlais DC, Perlis T, Arasteh K, et al. Reductions in hepatitis C virus and HIV infections among injecting drug users in New York City, 1990–2001. *AIDS* 2005; **19**: S20–S25.
- McNaughton AL, Cameron ID, Wignall-Fleming EB, et al. Spatiotemporal reconstruction of the introduction of hepatitis C virus into Scotland and its subsequent regional transmission. *J Virol* 2015; **89**: 11223–32.
- Prati D. Transmission of hepatitis C virus by blood transfusions and other medical procedures: a global review. *J Hepatol* 2006; **45**: 607–16.
- Drucker E, Alcabes PG, Marx PA. The injection century: massive unsterile injections and the emergence of human pathogens. *Lancet* 2001; **358**: 1989–92.
- Krause G, Trepka MJ, Whisenhunt RS, et al. Nosocomial transmission of hepatitis C virus associated with the use of multidose saline vials. *Infect Control* 2003; **24**: 122–27.
- Widell A, Christensson B, Wiebe T, et al. Epidemiologic and molecular investigation of outbreaks of hepatitis C virus infection on a pediatric oncology service. *Ann Intern Med* 1999; **130**: 130–04.
- Tallis GF, Ryan GM, Lambert SB, et al. Evidence of patient-to-patient transmission of hepatitis C virus through contaminated intravenous anaesthetic ampoules. *J Viral Hepat* 2003; **10**: 234–39.
- Petit JM, Bour JB, Aho LS, Castaneda A, Vaillant G, Brun JM. HCV and diabetes mellitus: influence of nosocomial transmission with the use of a finger stick device. *Am J Gastroenterol* 1999; **94**: 1709–10.
- Ross RS, Viazov S, Roggendorf M. Phylogenetic analysis indicates transmission of hepatitis C virus from an infected orthopedic surgeon to a patient. *J Med Virol* 2002; **66**: 461–67.
- Tanaka Y, Kurbanov F, Mano S, et al. Molecular tracing of the global hepatitis C virus epidemic predicts regional patterns of hepatocellular carcinoma mortality. *Gastroenterology* 2006; **130**: 703–14.
- Henquell C, Guglielmini J, Verbeeck J, et al. Evolutionary history of hepatitis C virus genotype 5a in France, a multicenter ANRS study. *Infect Genet Evol* 2011; **11**: 496503.
- Valdiserri R, Khalsa J, Dan C, et al. Confronting the emerging epidemic of HCV infection among young injection drug users. *Am J Public Health* 2014; **104**: 816–21.
- Hagan H, Jordan AE, Neurer J, Cleland CM. Incidence of sexually transmitted hepatitis C virus infection in HIV-positive men who have sex with men. *AIDS* 2015; **29**: 2335–45.