From prevalence to incidence - a new approach in the hospital setting

Niklas Willrich[1], Sebastian Haller[1], Tim Eckmanns[1], Benedikt Zacher[1], Tommi Kärki[2], Diamantis Plachouras[2], Alessandro Cassini[2], Carl Suetens[2], Michael Behnke[3], Petra Gastmeier[3], Jan Walter[1]

[1] Department for Infectious Disease Epidemiology, Robert Koch Institute (RKI), Berlin, Germany, [2] European Centre for Disease Prevention and Control (ECDC), Solna, Sweden, [3] Institute of Hygiene and Environmental Medicine, Charité University Medicine Berlin, Berlin, Germany.

Corresponding author:
Niklas Willrich
Robert Koch Institute,
Berlin
e-mail: WillrichN@rki.de

Keywords: Prevalence, incidence, healthcare-associated infections, point-prevalence surveys

Abstract

Point-prevalence surveys (PPSs) are often used to estimate the prevalence of healthcare-associated infections (HAIs). Methods for estimating incidence of HAIs from prevalence have been developed, but application of these methods is often difficult because key quantities, like the average length of infection, cannot be derived directly from the data available in a PPS. We propose a new theory-based method to estimate incidence from prevalence data dealing with these limitations and compare it to other estimation methods in a simulation study. In contrast to previous methods, our method does not depend on any assumptions on the underlying distributions of length of infection and length of stay. As a basis for the simulation study we use data from the second study of nosocomial infections in Germany (Nosokomiale Infektionen in Deutschland, Erfassung und Prävention - NIDEP2) and the European surveillance of HAIs in intensive care units (HAI-Net ICU). The new method compares favourably with the other estimation methods and has the advantage of being consistent in its behaviour across the different setups. It is implemented in an R-package `prevtoinc` which will be freely available on CRAN (http://cran.r-project.org/).

INTRODUCTION

Epidemiological information on healthcare-associated infections (HAIs) is often acquired by means of point-prevalence surveys (PPSs). Large-scale PPSs are regularly performed by the European Centre for Disease Prevention and Control (ECDC) (1, 2), as well as the US Centers for Disease Prevention and Control (CDC) (3, 4). While the prevalence of HAIs is an important measure in itself, epidemiologists are usually more interested in the incidence of HAIs. For example, estimations of the burden of HAIs often rely on incidence rather than prevalence (5). Therefore, methods of estimating the incidence rates from the data of PPSs are needed. Under general conditions, the incidence and prevalence can be estimated from one another (6). The question of estimating incidence from prevalence in the context of HAIs has been addressed in the 1980s by two articles (7, 8). The method developed by Rhame and Sudderth (7) is the most commonly applied method for estimating incidence from prevalence (1-3, 5, 9-14). This method however has several limitations:

The Rhame-Sudderth formula was developed using a definition of prevalence that included active and cured infections on the day of the PPS and that is different from the one usually applied in PPSs of HAIs. Another problem with the application of the formula is that it requires a method to estimate the average length of stay and the average length of infection based on data available on the day of the PPS. Without estimates of these quantities from other sources, the application of the estimation method is challenging, because usually only the data obtained on the day of the PPS are available.

In this article, we propose a novel approach dealing with these limitations of estimating incidence from prevalence of HAIs. The proposed approach uses state-of-the-art statistical techniques to estimate the average length of infection and average length of stay in the whole patient population from samples of lengths of infection and hospital stay up to the day of the PPS without relying on any assumption about the distributions of these quantities. We evaluated the new method by comparing it with existing procedures in the literature through simulation studies based on data from the second study of nosocomial infections in Germany (Nosokomiale Infektionen in Deutschland, Erfassung und Prävention - NIDEP2) (15) and from the European surveillance of HAIs in intensive care units (HAI-Net ICU) (16, 17), as well as theoretical distributions.

METHODS

Notation

In general, we used the variable X to indicate a randomly sampled duration from the whole population and L for a randomly sampled duration from the PPS (the duration for a randomly selected patient included in the PPS). L is expected to be on average larger than X, due to the phenomenon of length-biased sampling (8, 18). We used A for the observed duration up to a fixed time for a randomly selected patient at that time point. This was applied to the length of stay and the length of infection.

We used X, A and L when it was not important to distinguish between the length of stay and the length of infection from a theoretical perspective.

The different concepts for the durations are illustrated in Fig. 1. The notation used in this article is explained in Table 1.

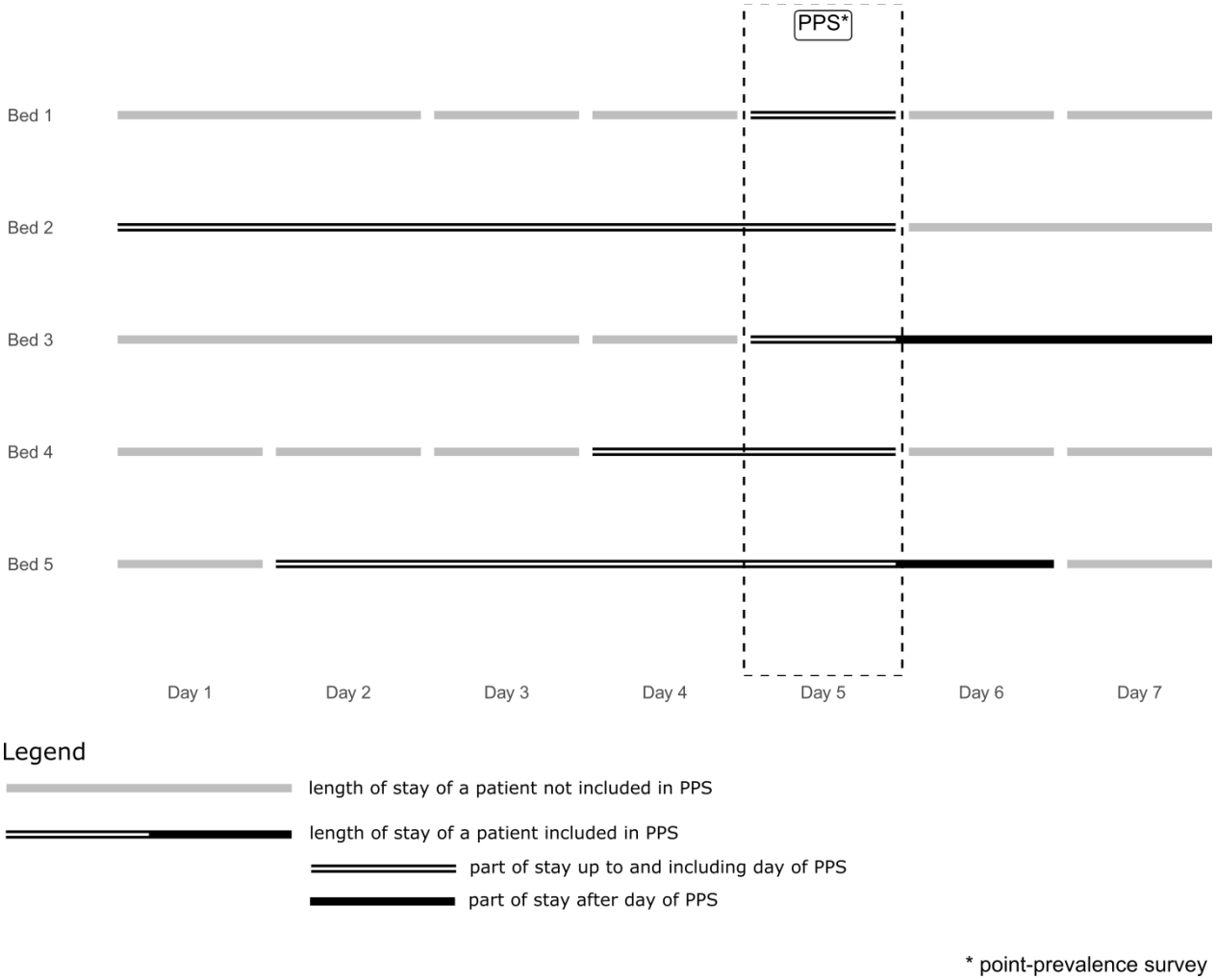

Fig. 1 *Illustration of different samplings for a hypothetical hospital: Line segments represent patients admitted in a hypothetical hospital. Sampling from X means selecting one of all the line segments at random, sampling from L means only sampling among the segments which intersect with the survey and sampling A means only sampling from the striped parts of these segments representing the part of stay up to and including the day of the PPS.*

Type of measure	Notation	Definition
Burden of disease		
Prevalence	$\boldsymbol{P}$	Expected proportion of patients with a healthcare-associated infection (HAI) on a fixed day
Prevalence according to Rhame and Sudderth	$\boldsymbol{P_{rhame}}$	Expected proportion of patients present on a fixed day who have or had a HAI during their hospital stay so far
Incidence rate	$\boldsymbol{I}$	Probability of an uninfected patient acquiring a HAI on a specific day (hazard rate)
Incidence proportion of HAIs per admission	$\boldsymbol{I_{pp}}$	Probability of a patient acquiring a HAI during one stay; similar to the concept introduced by Rhame and Sudderth (7)
Duration		
Length of stay	$\boldsymbol{L_{los}}$	Length of stay in hospital for a random patient included in the PPS
	$\boldsymbol{X_{los}}$	Length of stay in hospital for a random patient in the whole population
	$\boldsymbol{A_{los}}$	Length of stay in hospital up to day of PPS for a random patient included in the PPS
Length of infection	$\boldsymbol{L_{loi}}$	Length of HAI (during hospital stay) for a random infected patient included in the PPS
	$\boldsymbol{X_{loi}}$	Length of HAI (during hospital stay) for a random infected patient sampled from the whole population
	$\boldsymbol{A_{loi}}$	Length of HAI up to day of PPS for a random infected patient included in the PPS
Length of stay after onset of infection	$\boldsymbol{L_{LN-INT}}$	Length of stay after onset of first infection for a random patient who acquired a HAI included in the PPS
	$\boldsymbol{X_{LN-INT}}$	Length of stay after onset of first infection for a random patient who acquired a HAI
	$\boldsymbol{A_{LN-INT}}$	Length of stay after onset of first infection up to day of PPS for a random patient included in the PPS
Estimators	$\hat{\boldsymbol{P}}, \hat{\boldsymbol{I}}_{\boldsymbol{pp}}, \ldots$	Estimators of theoretical measures are identified by a ^
Population averages	$\boldsymbol{x_{los}}, \boldsymbol{a_{los}}, \boldsymbol{x_{LN-INT}}, \boldsymbol{a_{LN-INT}}, \ldots$	Population averages for the measures of length described above are denoted by the corresponding lowercase letter

Table 1 *Notation used in this article*

Rhame and Sudderth formula

In line with previous authors (7, 8), we assumed that the patient population is in steady state, i.e. the distribution of characteristics of our sample of patients does not depend on the specific day of the survey.

The original formula of Rhame and Sudderth (7) for the incidence per admission I_{pp} (slightly simplified and adapted to our notation) is:

$$I_{pp} = P_{rhame} \frac{x_{los}}{x_{LN-INT}},$$

where x_{los} denotes the average length of stay of a patient, x_{LN-INT} is the average length of stay for patients after they acquire their first HAI and I_{pp} is the estimate of the incidence per admission. In this original formulation, P_{rhame} is calculated by counting all patients who had at least one HAI *up to* the time of the survey (and not just the patients that have an active HAI on the date of the survey) and dividing by the total number of patients.

As pointed out above there are two points that complicate the application of this formula in this form:

(1) often the PPSs only count patients with *active* infections on the day of the PPS.

In these cases, theoretical considerations then require that the term x_{LN-INT} is replaced by a term x_{loi} which gives the average length of a HAI (see supplement S1).

(2) samples of X_{los}, X_{loi} (or X_{LN-INT}) are often not available and only the length of stay A_{los} and possible length of infection A_{loi} up to the day of the PPS are available.

New approach

To estimate the distributions of length of stay and length of infection from the observed lengths of stay up to the day of the PPS, we proceeded in two steps:

- We estimated the distributions of length of stay and length of infection up to the day of the PPS (in our notation A_{los} and A_{loi}) from the available data,
- We calculated from these distributions the expected lengths (x_{los} and x_{loi}) for the whole population.

For the first part, we used an estimator which ensures the monotonicity of the estimated distribution, because the distribution of A is always monotonously decreasing. This can be demonstrated by the timeline of occupancy of a hypothetical bed: on average there will always be more patients for whom it is the first day of their stay than the second day, more patients for whom it is the second day of their stay than the third and so on. We use a *Grenander* estimator for discrete distributions described and studied by Jankowski and Wellner (19). This estimator is the maximum likelihood estimator for a discrete monotonously decreasing distribution and is therefore a canonical choice. It is well-studied

from a theoretical point and has good properties like consistency and $\sqrt{n}$ - rate of convergence (19)

Following Freeman and Hutchison (8), in the steady state the relation between prevalence P and incidence rate I can be written as:

$$I = \frac{P}{(1-P)x_{loi}},$$

where x_{loi} is the average length of a HAI in the whole population. To get this equation into the form of Rhame and Sudderth (7), we multiply by the expected length of stay of random patients that are susceptible to an infection $(1-P)x_{los}$ and get

$$I_{pp} = P\frac{x_{los}}{x_{loi}}.$$

To express x_{loi} in terms of A_{loi}, we note the following formula with N_{pat} the total number of patients at the hospital on the survey day:

$$I(1-P)N_{pat} = \mathbb{P}(A_{loi}=1)\cdot PN_{pat}\ ,$$

where $(1-P)N_{pat}$ is the average number of patients at risk, $\mathbb{P}(A_{loi}=1)$ is the average proportion of patients with HAI on the first day of infection and PN_{pat} is the average number of patients with a HAI.

Both sides of the equation represent the number of average new infections per day; the left hand side as the incidence rate I per patient-day-at-risk times the number of patients at risk $(1-P)N_{pat}$ and the right hand side as the average number of HAI cases on the first day of infection. Therefore

$$I = \frac{P}{1-P}\mathbb{P}(A_{loi}=1),$$

where $\mathbb{P}(A_{loi}=1)$ is the probability of sampling $A_{loi}=1$. By comparing with the original incidence rate formula, this gives us, the simple relation

$$x_{loi} = \frac{1}{\mathbb{P}(A_{loi}=1)}.$$

An alternative, more formal route to the formula is based on renewal theory (8) and specifically Eqn. 2.16 from Haviv (20).

This leads to the estimator:

$$\hat{x}_{loi} = \frac{1}{\hat{\mathbb{P}}\,(A_{loi}=1)}.$$

with $\hat{\mathbb{P}}\,(A_{loi}=1)$ an estimator of $\mathbb{P}(A_{loi}=1)$. We call $\hat{x}_{\text{gren}}$ the estimator for x based on this procedure with the Grenander estimator (19) for $\hat{\mathbb{P}}\,(A_{loi}=1)$.

The general method is equally applicable for the estimation of x_{los}, x_{loi}, x_{LN-INT} and one can construct similar estimators. The derivation of the respective estimators is based on Eqn. 2.16 from Haviv (20).

Design of simulations

To assess the performance of our new estimator, we compared it in a simulation study to a selection of other estimators from the literature. Simulations were performed using R 3.5.1 (21) with the `prevtoinc` package which will be freely available on CRAN (http://cran.r-project.org/).

In a first step, we assessed the quality of the estimators for $x_{loi}, x_{LN-INT}, x_{los}$.

The setup was the following: a distribution for X_{loi} was chosen and the corresponding distributions for A_{loi} and L_{loi} were derived. A sample of n values from A_{loi} and L_{loi} was drawn and, based on this sample, all considered estimators of x_{loi} were calculated. We repeated this procedure m times and calculated the root-mean square deviation (RMSD) for each estimator.

An analogous procedure was used to benchmark estimators for x_{los} and x_{LN-INT}.

We performed repeated simulations to assess the performance of estimators for I based on simulated PPS data as follows: The number of patients n in the PPS was fixed, as well as a distribution for X_{loi} and a value P = 0.05 was fixed for the prevalence. For each patient, the presence of a HAI was determined by a sample from a Bernoulli distribution with as parameter P. In a next step, for patients with HAIs a joint sample of A_{loi} and L_{loi} was sampled from the chosen distribution. To assess the performance of the estimators for I_{pp} we additionally sampled A_{los} and L_{los} jointly for all patients. For a simulation distribution of X_{LN-INT}, assessment of estimators was performed in an analogous way replacing P by P_{rhame} = 0.2. For further parameters of the simulations see supplement S3.

Estimators for comparison

We used the following estimators to benchmark the performance of our new estimator.

- *pps.median* - estimator based on the median duration up to PPS (1)
$$\hat{\hat{x}}_{pps.median} = \text{median}(A),$$
where median(A) is the median of samples of the observed A,
- *pps.mean* - alternative estimator used in (1) based on the mean instead of the median
$$\hat{\hat{x}}_{pps.mean} = \text{mean}(A).$$
- *L.full* - estimator based on samples from the PPS with information on L based on the transformation formula $\hat{\hat{x}} = \frac{1}{\text{mean}(1/L)}$

The transformation formula uses the theoretical relationship between X and L derived in Eqn. 7 in (8). When comparing the performance of estimators, one has to keep in mind

that *L.full* uses the information on the whole durations L instead of only information on A .

The estimators can be used to estimate x_{loi}, x_{LN-INT} or x_{los} depending on which duration up to PPS A we use.

Mixed estimator

We also experimented with the combination of different estimators by weighting. As will be seen in the results section, for small samples the estimator *gren* has high variance. While it is unbiased (inside the model), it could be advantageous to combine it with a biased estimator with lower variance for small sample sizes. As a specific case, we introduced the following estimator *pps.mixed* based on the estimators *pps.mean* and *gren*:

$$\hat{x}_{pps.mixed} = \alpha(n)\,\hat{x}_{pps.mean} + (1-\alpha(n))\,\hat{x}_{gren}.$$

The function α is chosen as a sigmoid function: $\alpha(n) := \frac{\exp(0.01\cdot(n-500))}{1+\exp(0.01\cdot(n-500))}$. This gives a smooth transition between *pps.mean* and the new estimator *gren* with equal weighting $\alpha = 0.5$ on $n = 500$. Again this type of estimator can be used for the estimation of x_{loi}, x_{LN-INT} or x_{los}.

Constructing estimators for I and I_{pp}

We estimated the theoretical prevalence P by taking the observed prevalence $\hat{P}$ on the day of the PPS as an estimate. We constructed the incidence rate estimator in the general form:

$$\hat{I} = \frac{\hat{P}}{1-\hat{P}} * \frac{1}{\hat{x}_{loi}}$$

and for the incidence proportion per admission:

$$\hat{I}_{pp} = \frac{\hat{P}}{\hat{x}_{loi}}\hat{x}_{los},$$

where one uses any of the above estimators for $\hat{x}_{loi}$ and $\hat{x}_{los}$. A similar estimator could be built by plugging in the corresponding estimators in the original Rhame-Sudderth formula using P_{rhame} and x_{LN-INT}.

Simulation distributions for X_{loi}, X_{LN-INT} and X_{los}

We used three different distributions for X_{loi}: a geometric distribution shifted to start on 1 with mean 8, a Poisson distribution shifted to start on 1 with mean 8. We selected the two theoretical distributions, Poisson and geometric, to assess the flexibility of the estimators. We also used an empirical distribution of X_{loi} based on data from the NIDEP2 – study (15). In this study, incidence and prevalence of HAIs were measured on a daily basis in eight

German hospitals during two eight-week periods (see supplement S2 for a further description of the data).

For simulation of X_{LN-INT}, we used an empirical distribution of X_{LN-INT} based on the HAI-Net ICU data from 2015, which monitored date of onset of HAI and date of discharge of patients with an ICU-acquired HAI in 1 365 intensive care units (ICUs) from 11 European Union Member States (see supplement S2 for a further description of the data) (15, 16). No information on the end of the HAI was available, which is why we used X_{LN-INT} and the original version of the Rhame-Sudderth formula for this simulation example.

We show the resulting distributions of X_{loi} in Fig. 2 and the distribution for X_{LN-INT} in Fig. 3.

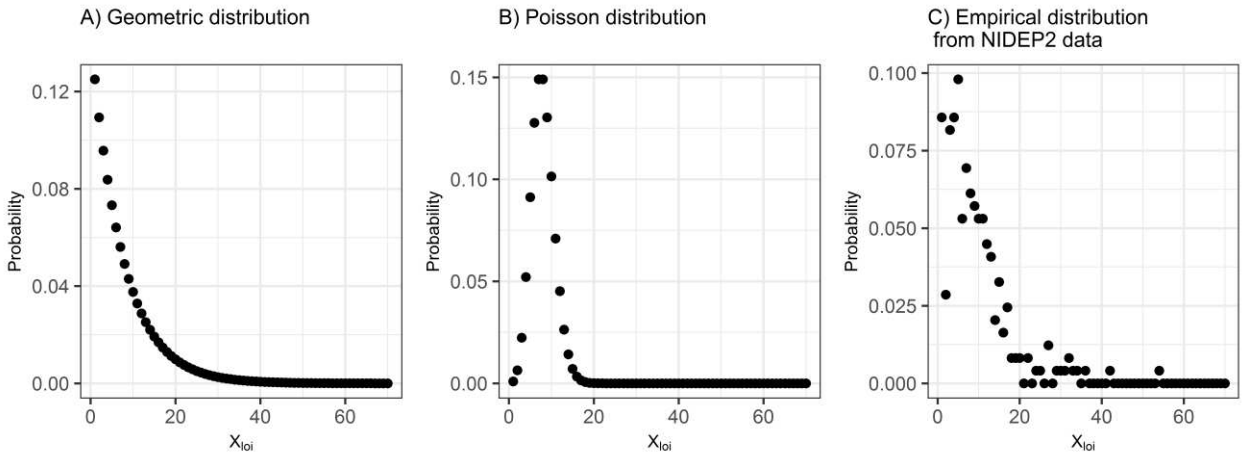

Fig. 2 *Distributions of X_{loi} for simulations*

Based on these distributions for X_{loi} we calculated the distributions of A_{loi} and L_{loi} (see Eqn. 2.14 and 2.16 (20) for the exact relation between these distributions).

For each simulation we then sampled n lengths of infection jointly from A_{loi} and L_{loi}. An analogous procedure was applied for sampling lengths of stay (A_{los} and L_{los}) and lengths of stay after infection (A_{LN-INT} and L_{LN-INT}). The distributions used for simulating the length of stay are shown in Fig. 3 and were based on data on lengths of stay from the NIDEP2-study (14) and HAI-Net ICU (15, 16).

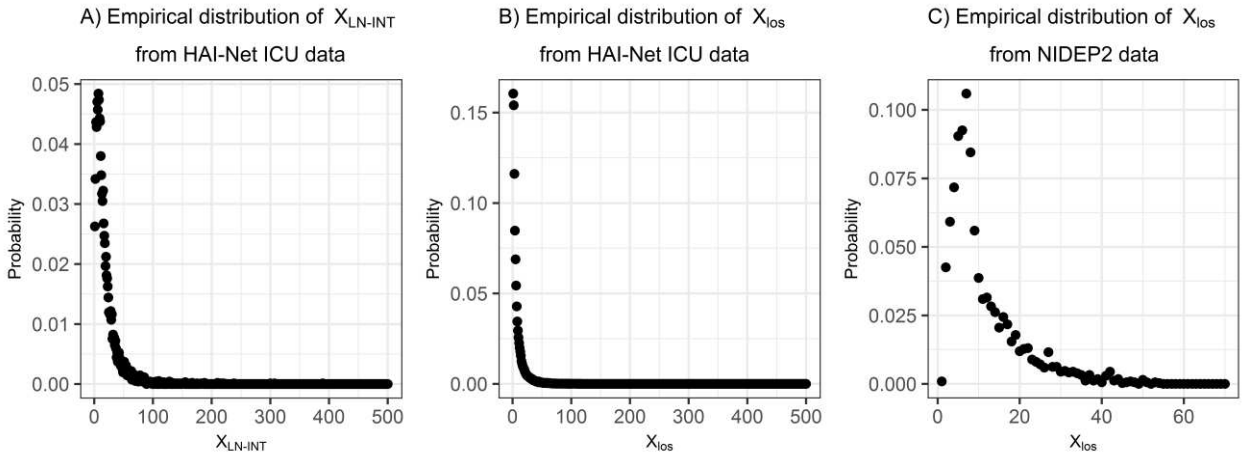

Fig. 3 *Empirical distribution of X_{LN-INT} from HAI-Net ICU data (A) and empirical distributions for X_{los} for NIDEP2 and HAI-Net ICU data (B and C)*

RESULTS

To assess the quality of the estimators, we measured the RMSD for increasing numbers of HAIs using different distributions (Fig. 4-7).

Simulations for x_{loi} and x_{LN-INT}

In Fig. 4, we present the RMSDs of the estimates of x_{loi}. We show the results for three examples of A_{loi} distributions. The simulations ranged from $n = 50$ to $n = 1000$. The estimators differed in the size of the RMSD, as well as in the convergence to zero along increasing sample sizes. In all three distributions, *pps.median* had the highest RMSD and generally did not converge to zero. The estimator pps.mean behaved similarly to *pps.median* in the case of the Poisson distribution. For the NIDEP2 distribution, it did not converge to zero, but stabilized on a lower RMSD compared to the Poisson distribution. In the case of the geometric distribution, *pps.mean* converged to zero with a low RMSD as could be expected for mathematical reasons (22), because $x_{loi} = a_{loi}$ for this specific distribution. *L.full* converged towards zero for all distributions and had among the lowest RMSD for all three settings. The RMSD of the new estimator *gren* converged towards zero in all three settings for large enough sample sizes and the magnitude of the RMSD was similar for all three distributions. The RMSD of *pps.mixed* for lower sample sizes was similar to the one of *pps.mean* and for larger sample sizes more like the new *gren* estimator as expected due to its construction.

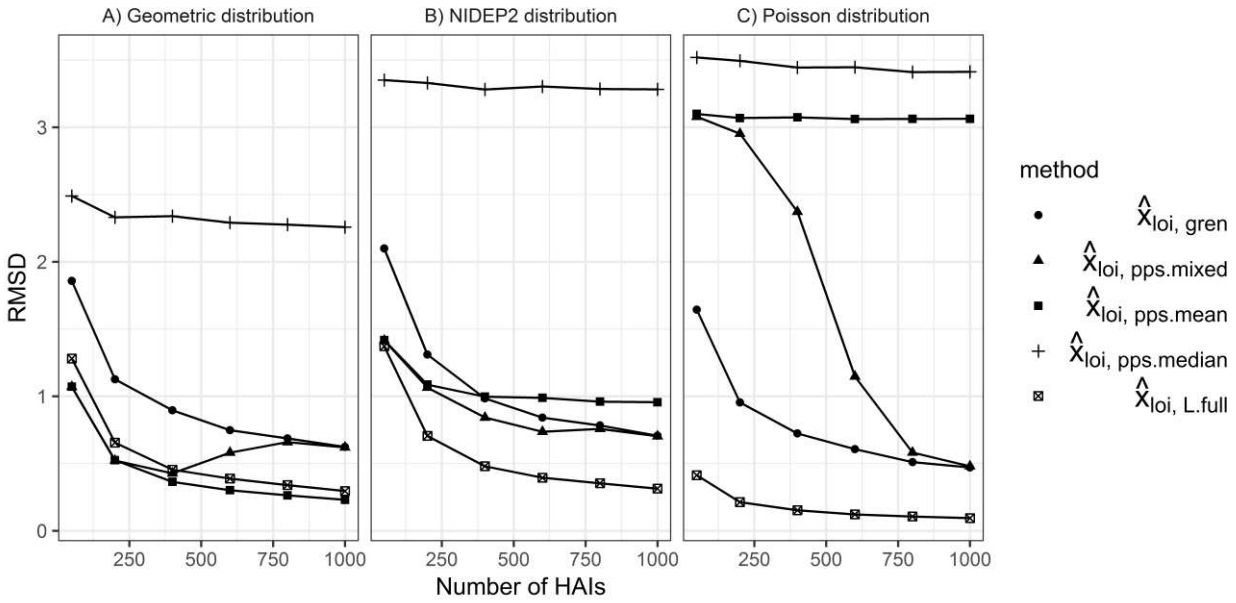

Fig. 4 Root-mean squared deviation (RMSD) of estimators of x_{loi} for 1000 simulations each along increasing size of samples of A_{loi} (resp. L_{loi} for L.full)

Similar plots for the bias (inside the model) and standard deviation can be found in the supplement S4 in Fig. S2 and S3. For boxplots of the estimators of $x_{loi}x_{lnint}$ and x_{los} see Fig. S4-S6 in supplement S5.

Results for the estimation of x_{LN-INT} for the HAI-Net ICU data are shown in Fig. 5. The simulations again ranged from $n = 50$ to $n = 1000$.
As previously, *pps.median* did not converge to zero and had the highest RMSD among the estimators. The estimator *pps.mean* stabilized at a significantly lower RMSD than *pps.median*, but did not converge towards zero either. *L.full* was again the best performing estimator in terms of RMSD. *gren* and *pps.mixed* behaved similarly to the estimation of x_{loi}. *gren* exhibited a lower RMSD than *pps.mean* as the sample size increased, and the RMSD behaviour of *pps.mixed* with increasing sample size was between that of *pps.mean* and *gren*.

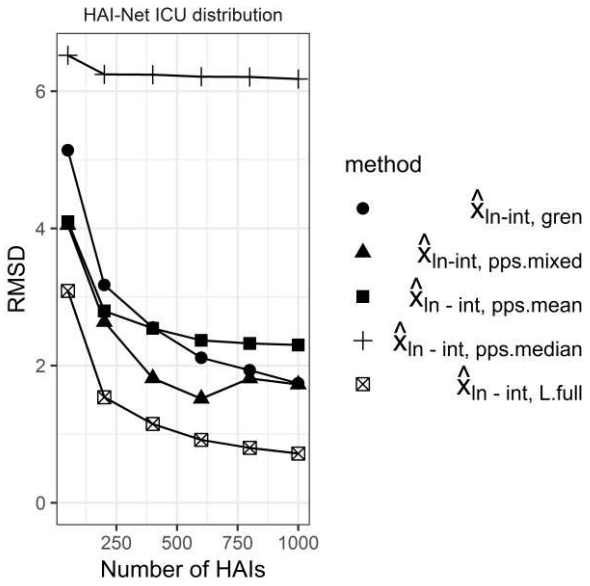

Fig. 5 *Root-mean squared deviation (RMSD) of estimators of x_{LN-INT} for 1000 simulations each along increasing size of samples of A_{LN-INT} (resp. L_{LN-INT} for L.full)*

Simulations for I

The results for the RMSDs of the estimators for I are shown in Fig. 6. In this figure the RMSD was divided by the theoretical incidence rate I to estimate the relative size of the error. The sample sizes ranged from $n = 500$ to $n = 20000$ patients in the simulated PPS. As expected, the RMSDs behave very similarly to the case of estimation of x_{loi} and x_{LN-INT} and the additional uncertainty in the estimation of P did not change the general patterns for the RMSDs shown in Fig. 4 and Fig. 5 when compared to Fig. 6.

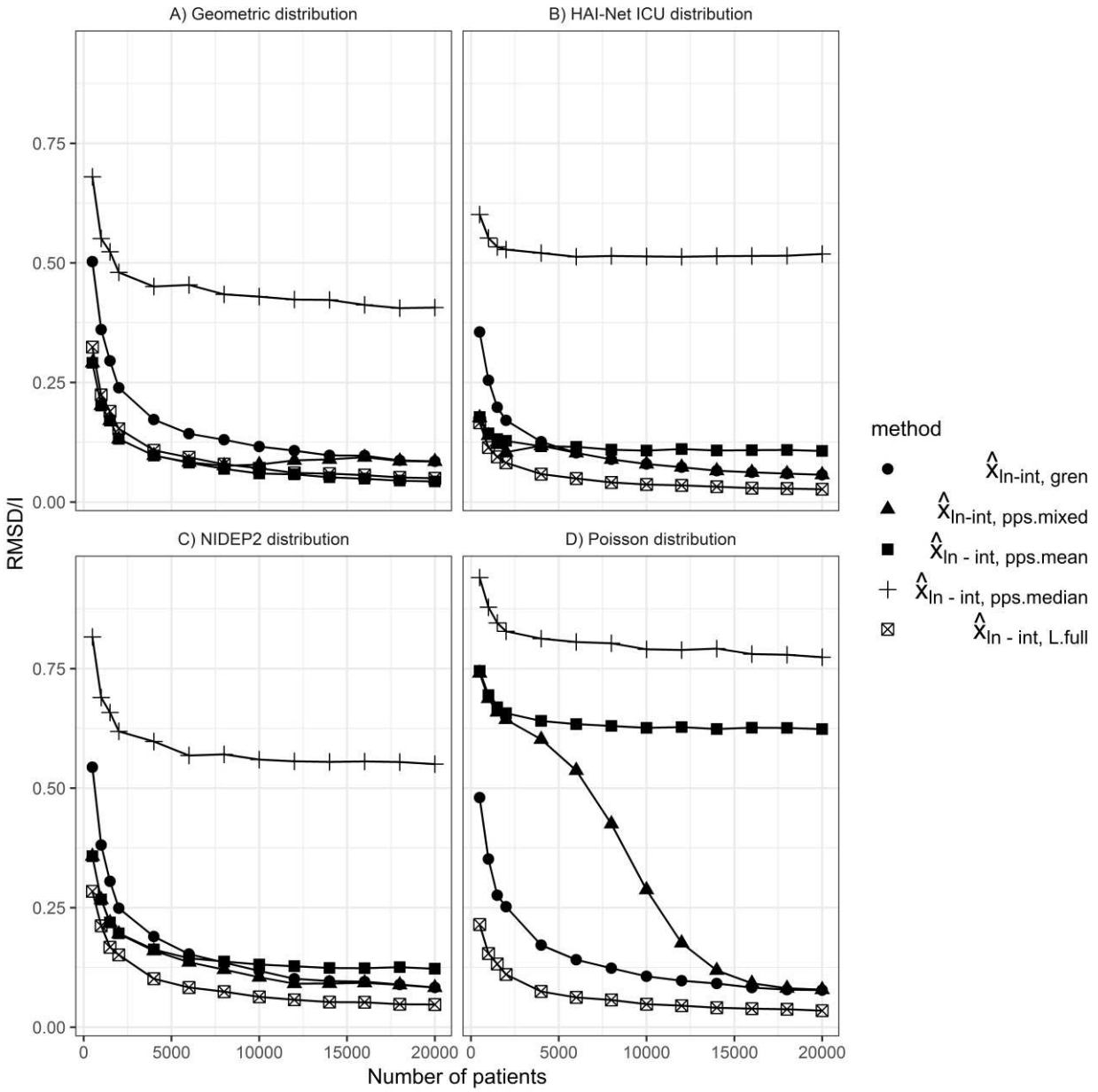

Fig. 6 *Root-mean squared deviation (RMSD) of estimators of I divided by theoretical incidence rate I along increasing size of samples from a simulated PPS based on 1000 simulations*

Simulations for x_{los}

Results for the length of stay in days are shown in Fig. 7. Again we presented the RMSD of the estimators of x_{los}. We used the empirical distributions of the length of stay from the NIDEP2 and HAI-Net ICU datasets.

For the NIDEP2 distribution of lengths of stay, *pps.median* again had the highest RMSD and did not converge toward zero. *pps.mean* did not converge toward zero either but stabilized at a lower RMSD than *pps.median*. *L.full* again had the lowest RMSD and *pps.mixed* and *gren* also had comparably low RMSD for larger samples ($n \geq 5000$).

For the HAI-Net ICU distribution of length of stay, the previous picture with respect to *pps.mean* and *pps.median* was reversed. *pps.mean* had the highest RMSD and did not converge towards zero, and *pps.median* had a lower RMSD but also did not converge towards zero. For *gren* and *L.full* the simulation results were similar to those obtained with the NIDEP2 distribution of length of stay. The estimator *pps.mixed* had a high RMSD compared to the other estimators for small sample sizes where the *pps.mean* component was dominant. For larger sample sizes, it behaved similarly to *L.full* and *gren*.

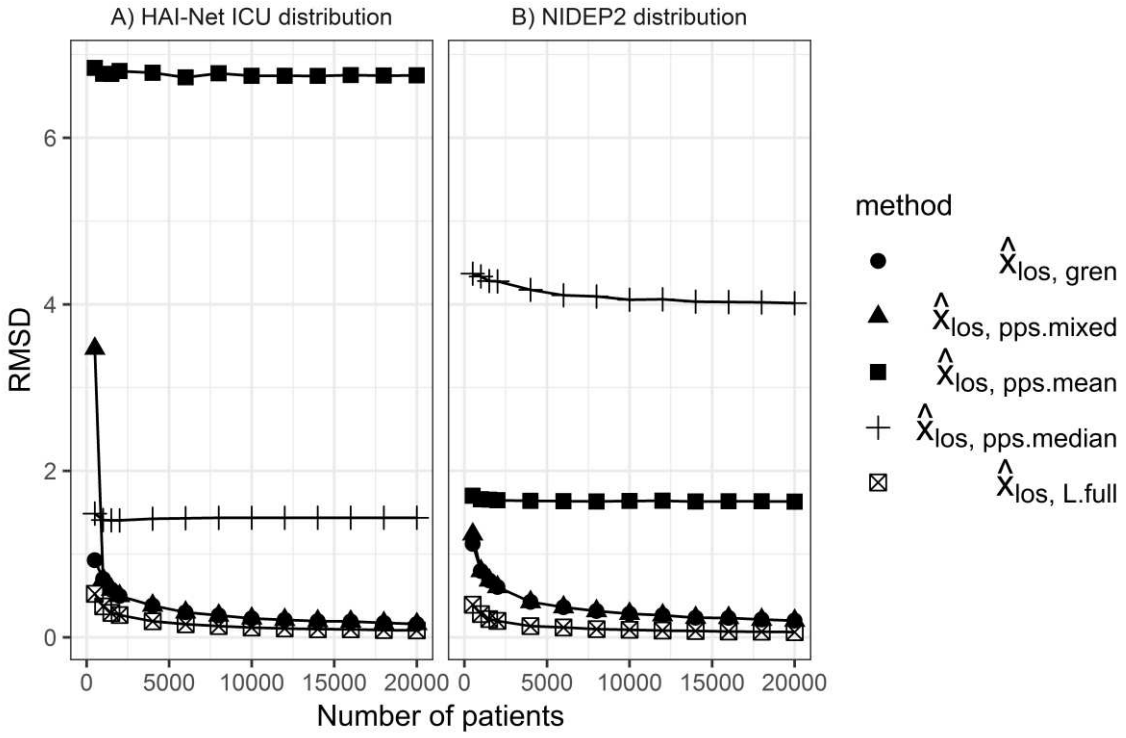

Fig. 7 *Root-mean squared deviation (RMSD) of estimators of x_{los} using 1000 simulations each along increasing size of samples of A_{los}*

Simulations for I_{pp}

For the incidence proportion of HAIs counted per admission, the RMSDs of the estimators are shown in Fig. 8 In this figure the RMSD was divided by the theoretical incidence proportion $I_{pp}$to estimate the relative size of the error. Almost all the estimators behaved very similarly in terms of RMSD. *pps.mean* and *pps.median* were the only estimators with a significantly higher RMSD than the other estimators for the HAI-Net ICU distribution. In the case of the NIDEP2 distribution and *pps.median*, the errors in the estimation of x_{loi} and x_{los}

seemed to cancel out almost exactly, reducing the RMSD for I_{pp} to levels comparable to the other estimators.

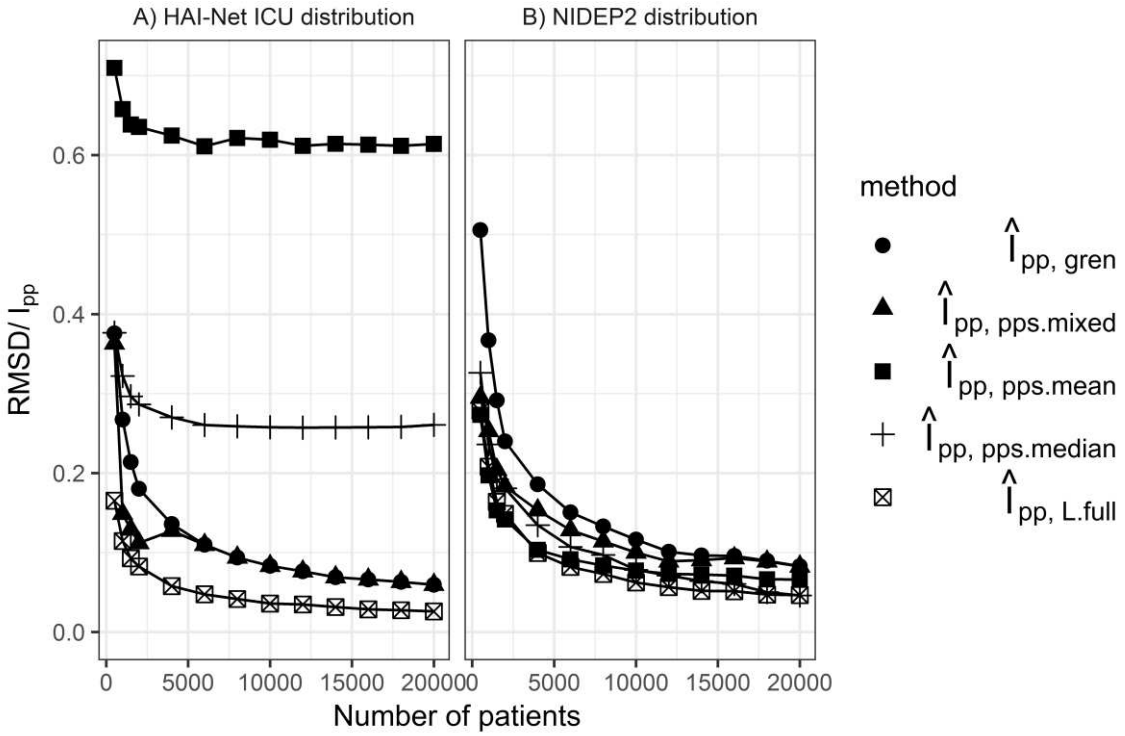

Fig. 8 *Root-mean squared deviation (RMSD) of estimators of I_{pp} divided by theoretical incidence proportion I_{pp} per admission along growing number of samples from a simulated PPS based on 1000 simulations*

DISCUSSION

We presented a method to estimate incidence from prevalence data available in a typical PPS setup. We used nonparametric estimators for length of stay and length of infection which exploit the monotonicity of A_{los} and A_{loi}. By means of a simulation study, we compared these estimators to other estimators that have been applied in previous studies.

The new *gren* estimator behaved consistently for different distributions, i.e. was more accurate with larger samples and in most cases was comparable to or better than the other estimators based on *A*. This was in contrast to *pps.median*, which generally did not converge to the true value. The estimator *pps.mean* did not perform overall as well as the *gren* estimator, but its variance for small sample sizes was lower. As expected, *L.full* performed better than or as well as all other estimators across all settings, but at the price of requiring knowledge of the full durations *L* which are typically not available in a PPS. We finally proposed the mixed estimator *pps.mixed* as a good compromise between the low variability

of the *pps.mean* for smaller samples and the consistent behaviour of the new estimator *gren* for larger samples. Altogether, the incidence estimate based on a PPS can be improved by more than 40% of the theoretical value (in terms of RMSD), compared to other estimators from the literature for a large enough PPS (see Fig. 8).

The new method presented in this article is a modification and update of the Rhame-Sudderth formula and is applicable in the setup of modern PPSs. The Rhame-Sudderth formula was published in the 1980s and, to our knowledge there have only been few methodological contribution addressing the questions of validity of the formula on a theoretical level since its publication. Mandel and Fluss (23) have proposed and studied incidence estimators, which generalize the Rhame-Sudderth,estimator but they depend on the use of the original Rhame-Sudderth prevalence definition and information about the total length of stay L_{los} for all patients in the survey. There have been attempts to evaluate the Rhame-Sudderth formula (14, 24), but these shared the limitation of using P instead of P_{rhame}, as intended in the original Rhame-Sudderth formula. Few studies distinguish between P and P_{rhame} and use the originally intended combination of prevalence and length of duration definitions. This often leads to the use of x_{LN-INT} as a proxy for length of infection x_{loi} (14, 24). It was suggested that x_{LN-INT} was not a good proxy for average length of infection (13) and instead some ad-hoc measure or external information could be used to estimate average length of infection (9, 11-13). Most of the articles remained critical of their own results. The ECDC-coordinated PPS of healthcare-associated infections and antimicrobial use in European acute care hospitals included data from over 200 000 patients across Europe and used the estimators *pps.mean* and *pps.median* (or more precisely a combination of these two) to estimate x_{LN-INT} stratified by participating country (1). Information on x_{los} was often obtained from external data sources. In the analysis of the latest ECDC-coordinated PPSs in acute care hospitals and long-term care facilities (2016-2017) (14) our proposed method has already been used for sensitivity analysis to compare with the estimator described above. European-level estimates were similar for the different estimators with few exceptions at individual country level.

The United States PPS coordinated by CDC (3) used stratification along factors thought to be predictive of the prevalence of HAIs. The estimators in (2) were based on medians of the durations-up-to-PPS similar to *pps.median* or external information and in (4) the original Rhame-Sudderth formula was used with the definition of prevalence P instead of P_{rhame} and with a length-biased version of x_{LN-INT} (i. e. l_{LN-INT}), and a length-biased version of x_{los} (i. e. l_{los}).

A main strength of our method is that we do not make any assumptions on the distributions of X_{loi} and X_{los}. Usually in a PPS we do not know the distribution of X_{loi} and X_{los}. This means that one criterion for selecting an estimator of I or I_{pp} is that it should behave well irrespective of the form of the unknown distribution. This is a criterion which, among the estimators using only duration-up-to-PPS information, was only fulfilled be the proposed estimator *gren* and the mixed estimator *pps.mixed* for larger samples ($n \geq 500$). This is supported not only by simulations, but also by theoretical considerations. Using only simulation studies to assess an estimator can be a source of error if the distributions on which the estimators are assessed differ significantly from the underlying distributions encountered in PPSs.

Our method has limitations. One is the requirement of a sufficiently large sample size to get an acceptable estimate. We took the sample size of 500 HAIs as a rule-of-thumb lower limit. It may be applied to smaller samples, but with a risk of lower precision. For a single medium-size hospital, repeated PPSs with aggregation of the results would need to be performed to reliably estimate the incidence of HAIs. Another limitation of our setup is that we counted multiple simultaneous or partially overlapping HAIs as one HAI. However, these in reality only comprise a small fraction of HAIs (1, 15) and therefore can be neglected. In addition, many of the limitations mentioned in the original article by Rhame and Sudderth also apply in this updated version, in particular the lack of explicit representation of outbreaks and the assumption that the risk that a patient acquires a HAI is independent of other patients' status. The new estimators *gren* and *pps.mixed* applied to the length of stay are sensitive to week day patterns in admissions and discharges (data not shown). Typically, for larger PPSs, data collection takes place on different weekdays for different hospitals or even different wards in the same hospital (1), thus mitigating the influence of these patterns on the estimates. Another issue is that patients on their first day of admission are sometimes underrepresented due to the PPS protocol, when e. g. only patients admitted before a fixed time are included in the PPS. The new estimators are based on the monotonicity assumption for the distribution of A_{los}, which is violated in this situation. One solution can be to let A denote full days of hospital stay and ignore the patients admitted on the date of the survey for the estimates of average length of stay, but include them in the estimate of the prevalence. Similar problems appear to a lesser extent for the first day of HAI. Other factors that need to be taken into account include the consistency of the application of case definitions for HAIs , and the representativeness of the hospital sample.

In conclusion, the proposed *gren* estimator and the combined estimator *pps.mixed* provide better estimates of the length of infection across a range of simulation settings when compared to previously used estimators and, in contrast to these, are grounded in theory. The simulations also serve as a guide of the sample size to include in a PPS required to estimate incidence. The method is shared and easily applicable with the help of the R package `prevtoinc`.

Abbreviations

CDC – Centers for Disease Control and Prevention
ECDC – European for Disease Control and Prevention
HAI – healthcare-associated infection
HAI-Net ICU – European surveillance of HAIs in intensive care units
NIDEP2 – second study of nosocomial infections in Germany (Nosokomiale Infektionen in Deutschland, Erfassung und Prävention)
PPS – point-prevalence survey
RMSD – root mean squared deviation

References

1. European Centre for Disease Prevention and Control. Point prevalence survey of healthcare-associated infections and antimicrobial use in European acute care hospitals. ECDC, Stockholm. 2013. http://ecdc.europa.eu/en/publications/Publications/healthcare-associated-infections-antimicrobial-use-PPS.pdf.
2. Suetens C, Latour K, Karki T, et al. Prevalence of healthcare-associated infections, estimated incidence and composite antimicrobial resistance index in acute care hospitals and long-term care facilities: results from two European point prevalence surveys, 2016 to 2017. Euro Surveill. 2018;23(46). doi:10.2807/1560-7917.ES.2018.23.46.1800516
3. Magill SS, Edwards JR, Bamberg W, et al. Multistate point-prevalence survey of health care-associated infections. N Engl J Med. 2014;370(13):1198-208. doi:10.1056/NEJMoa1306801
4. Magill SS, O'Leary E, Janelle SJ, et al. Changes in Prevalence of Health Care-Associated Infections in U.S. Hospitals. N Engl J Med. 2018;379(18):1732-44. doi:10.1056/NEJMoa1801550
5. Cassini A, Plachouras D, Eckmanns T, et al. Burden of Six Healthcare-Associated Infections on European Population Health: Estimating Incidence-Based Disability-Adjusted Life Years through a Population Prevalence-Based Modelling Study. PLoS Med. 2016;13(10):e1002150. doi:10.1371/journal.pmed.1002150
6. Rothman KJ, Greenland S, Lash TL. Modern epidemiology. 3rd ed. ed. Philadelphia, Pa.: Lippincott Williams & Wilkins; 2008.
7. Rhame FS, Sudderth WD. Incidence and prevalence as used in the analysis of the occurrence of nosocomial infections. Am J Epidemiol. 1981;113(1):1-11.
8. Freeman J, Hutchison GB. Prevalence, incidence and duration. Am J Epidemiol. 1980;112(5):707-23.
9. Berthelot P, Garnier M, Fascia P, et al. Conversion of prevalence survey data on nosocomial infections to incidence estimates: a simplified tool for surveillance? Infect Control Hosp Epidemiol. 2007;28(5):633-6. doi:10.1086/513536
10. Gastmeier P, Brauer H, Sohr D, et al. Converting incidence and prevalence data of nosocomial infections: results from eight hospitals. Infect Control Hosp Epidemiol. 2001;22(1):31-4. doi:10.1086/501821

11. Graves N, Nicholls TM, Wong CG, Morris AJ. The prevalence and estimates of the cumulative incidence of hospital-acquired infections among patients admitted to Auckland District Health Board Hospitals in New Zealand. Infect Control Hosp Epidemiol. 2003;24(1):56-61. doi:10.1086/502116
12. Kanerva M, Ollgren J, Virtanen MJ, Lyytikainen O, Prevalence Survey Study G. Estimating the annual burden of health care-associated infections in Finnish adult acute care hospitals. Am J Infect Control. 2009;37(3):227-30. doi:10.1016/j.ajic.2008.07.004
13. King C, Aylin P, Holmes A. Converting incidence and prevalence data: an update to the rule. Infect Control Hosp Epidemiol. 2014;35(11):1432-3. doi:10.1086/678435
14. Meijs AP, Ferreira JA, SC DEG, Vos MC, Koek MB. Incidence of surgical site infections cannot be derived reliably from point prevalence survey data in Dutch hospitals. Epidemiol Infect. 2017;145(5):970-80. doi:10.1017/S0950268816003162
15. Gastmeier P, Brauer H, Forster D, Dietz E, Daschner F, Ruden H. A quality management project in 8 selected hospitals to reduce nosocomial infections: a prospective, controlled study. Infect Control Hosp Epidemiol. 2002;23(2):91-7. doi:10.1086/502013
16. European Centre for Disease Prevention and Control. European surveillance of healthcare-associated infections in intensive care units – HAI-Net ICU protocol, version 1.02. Stockholm. 2015. http://ecdc.europa.eu/en/publications/Publications/healthcare-associated-infections-HAI-ICU-protocol.pdf
17. European Centre for Disease Prevention and Control. Healthcare-associated infections acquired in intensive care units. In: ECDC. Annual epidemiological report for 2015. ECDC, Stockholm. 2017. https://ecdc.europa.eu/sites/portal/files/documents/AER_for_2015-healthcare-associated-infections_0.pdf.
18. Arratia R, Goldstein L, Kochman F. Size bias for one and all. In: ArXiv E-Prints. 2013. https://arxiv.org/abs/1308.2729.
19. Jankowski HK, Wellner JA. Estimation of a discrete monotone distribution. Electron J Stat. 2009;3:1567-605. doi:10.1214/09-EJS526
20. Haviv M. Queues - a course in queuing theory. New York: Springer; 2013.
21. R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria; 2018.
22. Freeman J, McGowan JE, Jr. Day-specific incidence of nosocomial infection estimated from a prevalence survey. Am J Epidemiol. 1981;114(6):888-901.
23. Mandel M, Fluss R. Nonparametric estimation of the probability of illness in the illness-death model under cross-sectional sampling. Biometrika. 2009;96(4):861-72. doi:10.1093/biomet/asp046
24. Rossello-Urgell J, Rodriguez-Pla A. Behavior of cross-sectional surveys in the hospital setting: a simulation model. Infect Control Hosp Epidemiol. 2005;26(4):362-8. doi:10.1086/502553

Supplement: From prevalence to incidence - a new approach in the hospital setting

S1: Mathematical model and technical details

Derivation of the conversion formula

The theoretical model used is that of a discrete renewal process (see Chapter 2, [1] for an introduction to renewal theory). Similar to Rhame and Sudderth [2] we see a single bed as the basic unit to be simulated. Beds are assumed to be occupied sequentially by patients, which can develop HAIs on each day of stay. The evolution of patients/infections per bed are assumed to be statistically independent. We assume that time t is progressing in patient-days.

We use the framework of an alternating renewal process between HAI / non-HAI. Theorem 4.8 from [3] gives us the relation

$$P = \frac{\mathbb{E}(X_{loi})}{\mathbb{E}(X_{\text{no infection}}) + \mathbb{E}(X_{loi})},$$

where $X_{\text{no infection}}$ is the length of occupation of a bed without any infection (the time in between two infections) and $\mathbb{E}$ will denote the expected value in the following . In our model we assume that the probability of acquiring a HAI on a given day is I and therefore one can see that in our model $X_{\text{no infection}}$ follows a geometric distribution with parameter I. This means $\mathbb{E}(X_{\text{no infection}}) = 1/I$. It follows:

$$P = \frac{\mathbb{E}(X_{loi})}{\frac{1}{I} + \mathbb{E}(X_{loi})}$$

After rearrangement this gives:

$$I = \frac{P}{1-P}\frac{1}{\mathbb{E}(X_{loi})} = \frac{P}{1-P}\frac{1}{x_{loi}}.$$

To get the version per patient we first introduce the notation $X_{\text{los w/o HAI}}$ for the length of stay of a patient only counting days without a HAI. We can then write:

$$I_{pp} = \mathbb{E}(1-(1-I)^{X_{\text{los w/o HAI}}}) \approx I\mathbb{E}(X_{\text{los w/o HAI}}),$$

where the approximation is valid if $I\mathbb{E}(X_{\text{los w/o HAI}})$ is small.

Finally, we note under the assumption that X_{los} is independent of whether or not an infection occured:

$$\mathbb{E}(X_{\text{los w/o HAI}}) = \mathbb{E}(X_{los}(1-P)),$$

as $(1-P)$ represents the average proportion of time of stay without a HAI present.

A proof of this is based on the following equalities, where for the distribution of limits Slutsky's theorem is used:

$$\mathbb{E}(X_{\text{los w/o HAI}}) = \lim_{t\to\infty} \frac{N_{\text{days w/o HAI}}(t)}{N_{pat}(t)} = \lim_{t\to\infty} \frac{N_{\text{days w/o HAI}}(t)}{t} \lim_{t\to\infty} \frac{t}{N_{pat}(t)} = (1-P)\cdot\mathbb{E}(X_{los}),$$

where $N_{pat}(t)$ is the number of patients in the bed up to time t and $N_{\text{days w/o HAI}}(t)$ denotes the number of patients days without HAI in the bed up to time t.

Putting together the above results gives the final conversion formula:

$$I_{pp} = \frac{P}{\mathbb{E}(X_{loi})}\mathbb{E}(X_{los}) = \frac{P}{x_{loi}}x_{los}.$$

All the theoretical arguments can be repeated by replacing X_{loi} by X_{LN-INT} and this would lead to the original Rhame-Sudderth formula. We note that P is then defined as the proportion of time a bed is occupied by a person who had or has a HAI, which is the definition of P_{rhame} from the main text.

Asymptotics of *gren* estimators

In our setting the *gren* estimators of the average durations will exhibit asymptotic normality. The proofs are based on the delta method and Prop. 3.4 and Prop. 3.6 from [4]. One gets

$$\sqrt{n}\left(\hat{x}_{gren} - x_{loi}\right) \underset{n\to\infty}{\Rightarrow} \mathcal{N}(0, \Sigma_{gren})$$

with $\Sigma_{gren} = \frac{x_{loi}^3}{1-1/x_{loi}}$ where $\Rightarrow$ denotes convergence in distribution, n the sample size and $\mathcal{N}(\mu, \Sigma)$ denotes a normal distribution with mean μ and covariance Σ.

As pointed out in [4] the asymptotics of the Grenander estimator for a discrete distribution are the same as for one based on the empirical estimator, i. e. taking just the empirically observed proportions as an estimate for the distribution of A. To visualize the relation between asymptotic results and simulations, we take x_{loi} as an example. We call the estimator based on the empirical proportions $\hat{x}_{loi,empirical}$. In Fig. S1 we compare the simulated standard deviations for $\hat{x}_{loi,gren}$, $\hat{x}_{loi,empirical}$ and the asymptotical approximation of the standard deviation $\sqrt{\Sigma_{gren}}$. As predicted by theory, the standard deviations of the estimators approach the asymptotic value and also $\hat{x}_{loi,gren}$ systematically has the lowest standard deviation.

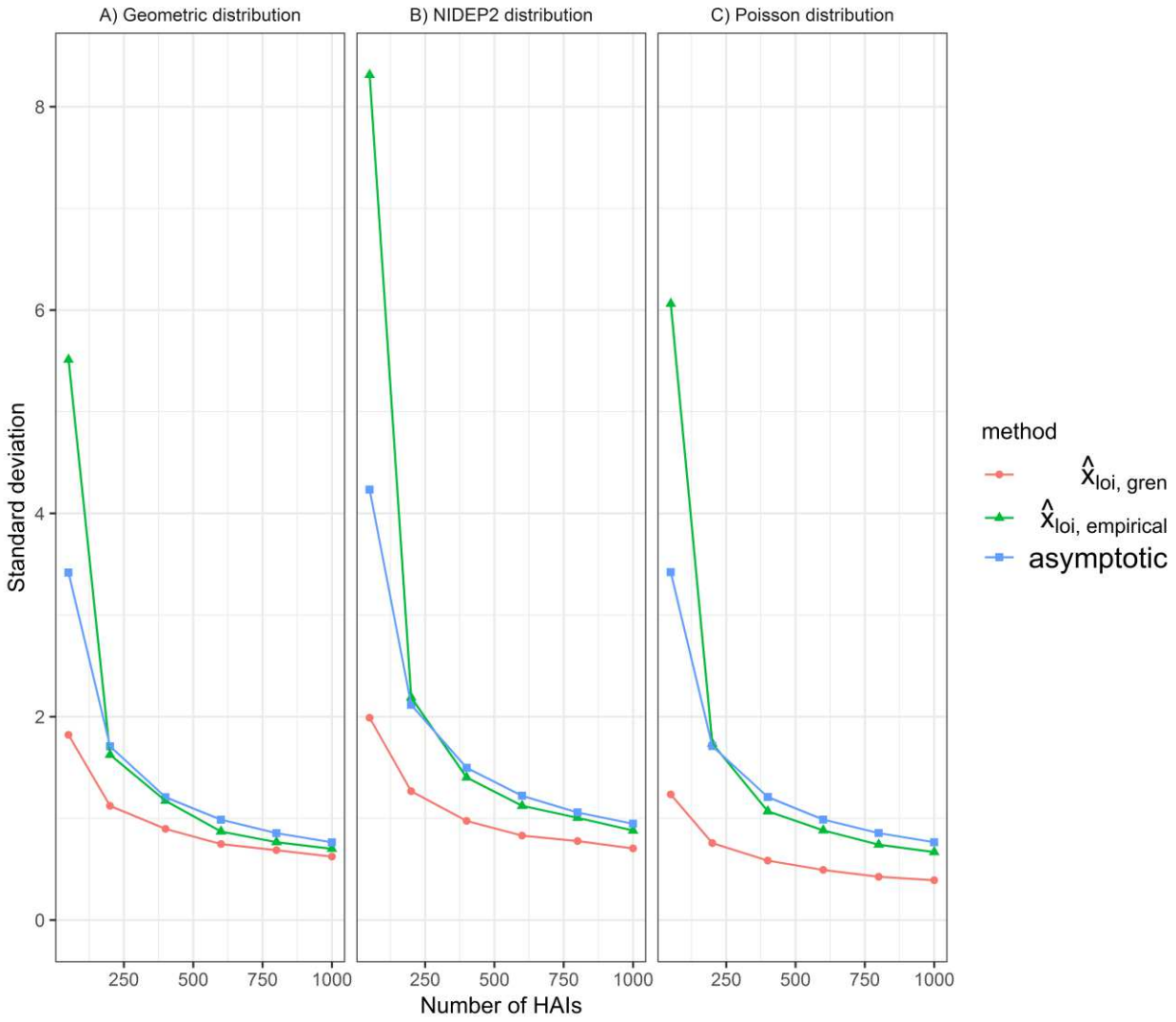

Fig. S1 *Standard deviation of estimators of for* $\hat{x}_{loi,empirical}$ *and* $\hat{x}_{loi,gren}$ *based on 1000 simulations each along growing number of samples of* A_{loi} *compared to approximation of standard deviation based on asymptotics (in green)*

S2: Description and characteristics of data sources

NIDEP2

As a data source for our simulations we used the NIDEP2 study [5]. We used data from eight German hospitals, where incidence and prevalence of HAIs were measured on a daily basis during two eight-week periods. The second monitoring period began after a randomly assigned intervention to four of the hospitals in which a range of additional infection prevention and control measures was introduced. A total of 7568 patients participated in the NIDEP2 study, among whom 487 had at least one HAI. We counted multiple simultaneous or partially overlapping periods of HAI as one HAI. For the generation of the

empirical NIDEP2-distribution of X_{loi}, we used the incidence data up to day 30 ($n_{hai} = 245$) for each period (as this corresponded to the point where the influence of the cutoff of the measurement after each incidence period was not discernible anymore). Infections already present on the first day of each period were excluded because there was no information on their duration before the start of the surveillance period.

HAI-Net ICU

As a second data source we used data from the European surveillance of HAIs in intensive care units (HAI-Net ICU) [6,7]. In 2015, a total of 141 955 patients from 1 365 intensive care units (ICUs) from 11 European Union Member States were included. Among these patients, 11 788 developed at least one HAI during their stay [7]. In the HAI-Net ICU dataset, the onset of the HAI and the date of discharge were available, but not any information on the end of the HAI, so we calculated X_{LN-INT} and used the original version of the Rhame-Sudderth formula to calculate estimates of x_{LN-INT}. This meant that we also used $Prhame$ as the measure of prevalence as we could not estimate P without the information at the end of the HAI. Based on the ICU data, P_{rhame} was around 22% for most of the year with a dip around the beginning/end of the year, which could be due to reporting practices. A further modification was to discount the first two days of stay in the ICU. According to the HAI-Net ICU protocol, infections are only considered ICU-acquired if they develop at least 48 hours after admission to the ICU [6]. All the estimators were modified accordingly.

S3: Specification of simulation parameters

We measured the performance of the methods by repeating the simulation 1000 times for different values of the sample size n. With this we estimated the bias (inside the model), standard deviation and RMSD of the estimators for x_{loi} and for the case of the HAI-Net ICU distribution x_{LN-INT}. The theoretical values for x_{loi} were the following: $x_{loi,geom} = 8$ for the geometric distribution, $x_{loi,nidep2} = 9.28$ and $x_{loi,pois} = 8$. In a next step, we assessed the performance of the methods for estimating I. We used a setup with a theoretical $P = 0.05$. This meant that each simulated patient had a probability of 0.05 having an active HAI on the day of the PPS. Taking this theoretical prevalence as given, we generated simulated PPS data with associated L_{loi} and A_{loi} for each patient with a HAI. For the HAI-Net ICU data, the same simulations were performed with X_{LN-INT} and $P_{rhame} = 0.20$ in place of X_{loi} and P. The average number of days from first HAI to discharge in the distribution was $x_{LN-INT,icu} = 18.38$. The theoretical incidence rate in the simulated models was $I = 0.0066$ (6.6 HAIs per 1000 patient-days at risk) for the Poisson and geometric distribution, $I = 0.0066$ (6.6 HAIs per 1000 patient-days at risk), $I = 0.0057$ (5.7 HAIs per 1000 patient-days at risk) for the NIDEP2 distribution and $I = 0.0136$ (13.6 HAIs per 1000 patient-days at risk) for the HAI-Net ICU distribution. These incidence figures should not be confused with the incidence that could be calculated from the original data sets. The results that we used were based on a fictitious fixed prevalence used for the simulations. We repeated these simulations along increasing sample sizes (numbers of patients) n with 500 simulations for each n to estimate the RMSD. To assess the quality of estimation of the length of stays, we also performed a simulation based on the empirical distribution of

lengths of stay in the NIDEP2 and HAI-Net ICU datasets. The procedure was analogous to the case of the length of infection. The mean values of the length of stay for the two distributions were: $x_{los,icu} = 8.44$ (discounting the first two days) and $x_{los,nidep2} = 11.01$. Finally, we assessed the estimators for I_{pp} by combining the distributions of length of stay and length of infection (or length of stay after infection) based on the NIDEP2 and HAI-Net ICU data under the assumption that the product of the marginal distributions of these quantities is a good approximation of the joint distribution. We used the same parameters and simulation sizes as for the estimation of I; the theoretical $I_{pp} = 0.062$ for the simulations from NIDEP2 data and $I_{pp} = 0.027$ from HAI-Net ICU data.

S4: Bias and standard deviation of the estimators for x_{loi}

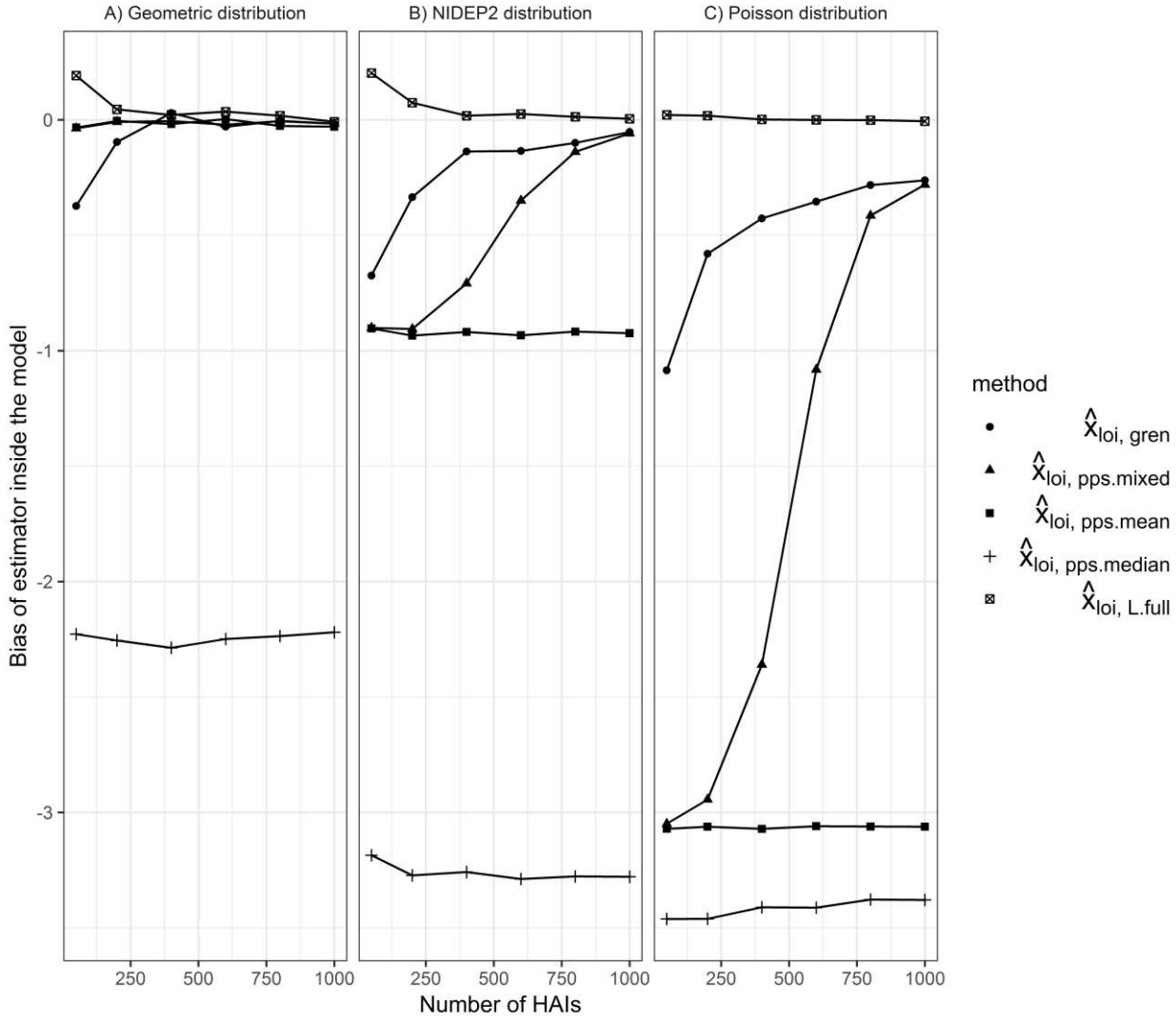

Fig. S2 *Bias of estimators (inside the model) of x_{loi} for 1000 simulations each along growing number of samples of A_{loi}*

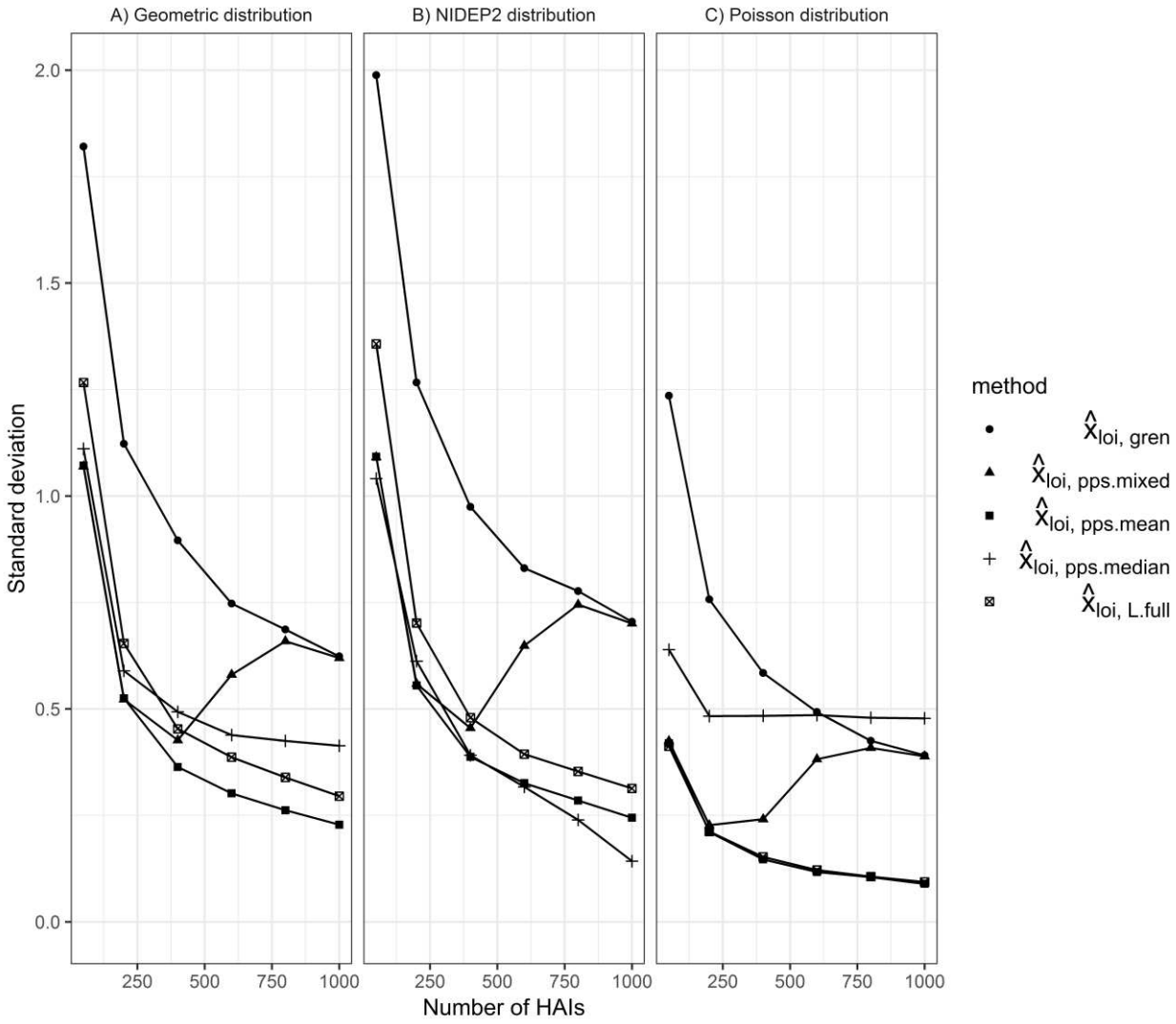

Fig. S3 *Standard deviation of estimators of x_{loi} for 1000 simulations each along growing number of samples of A_{loi}*

S5: Boxplots for simulations of estimates for x_{loi}, x_{LN-INT} and x_{los}

In this section, we present boxplots of the estimates for x_{loi}, x_{LN-INT} and x_{los}. The boxplots are specified in the following way: The lower and upper hinges correspond to the first and third quartiles. The upper whisker extends from the hinge to the largest value no further than 1.5 · IQR away from the hinge (IQR being the inter-quartile range). The lower whisker extends from the hinge to the smallest value at most 1.5 · IQR away from the hinge. Data points beyond the end of the whiskers are plotted individually.

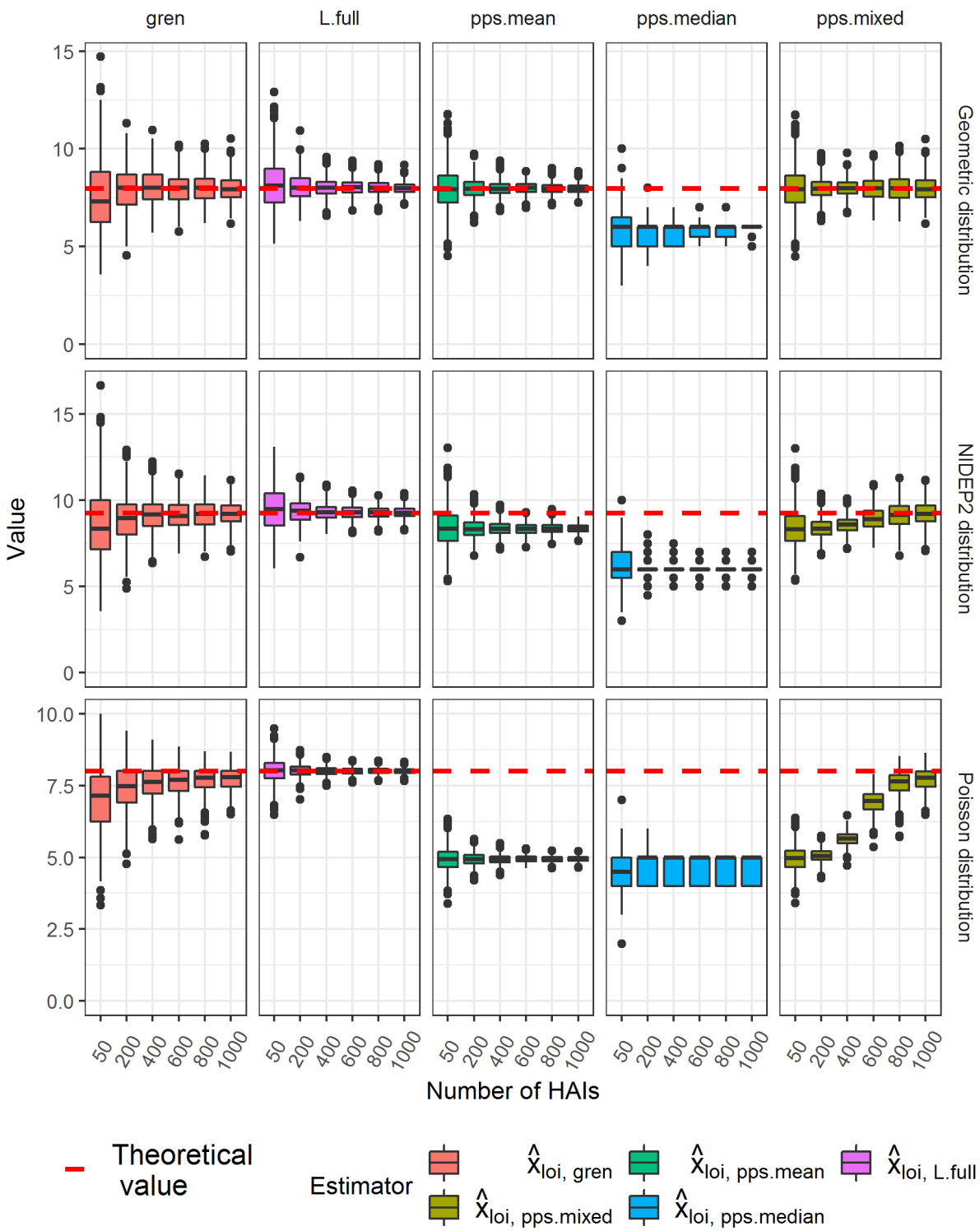

Fig. S4 *Boxplots of estimates of x_{loi} for 1000 simulations each along growing number of samples of A_{loi}*

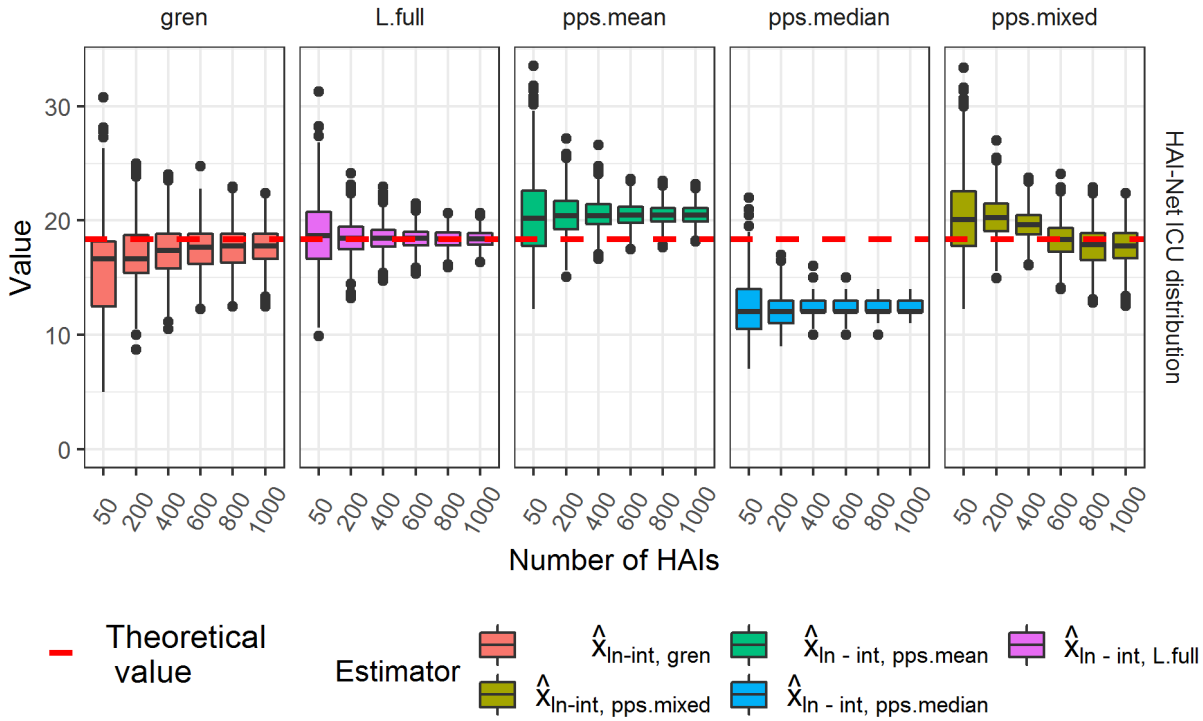

Fig. S5 *Boxplots of estimates of x_{LN-INT} for 1000 simulations each along growing number of samples of A_{LN-INT}*

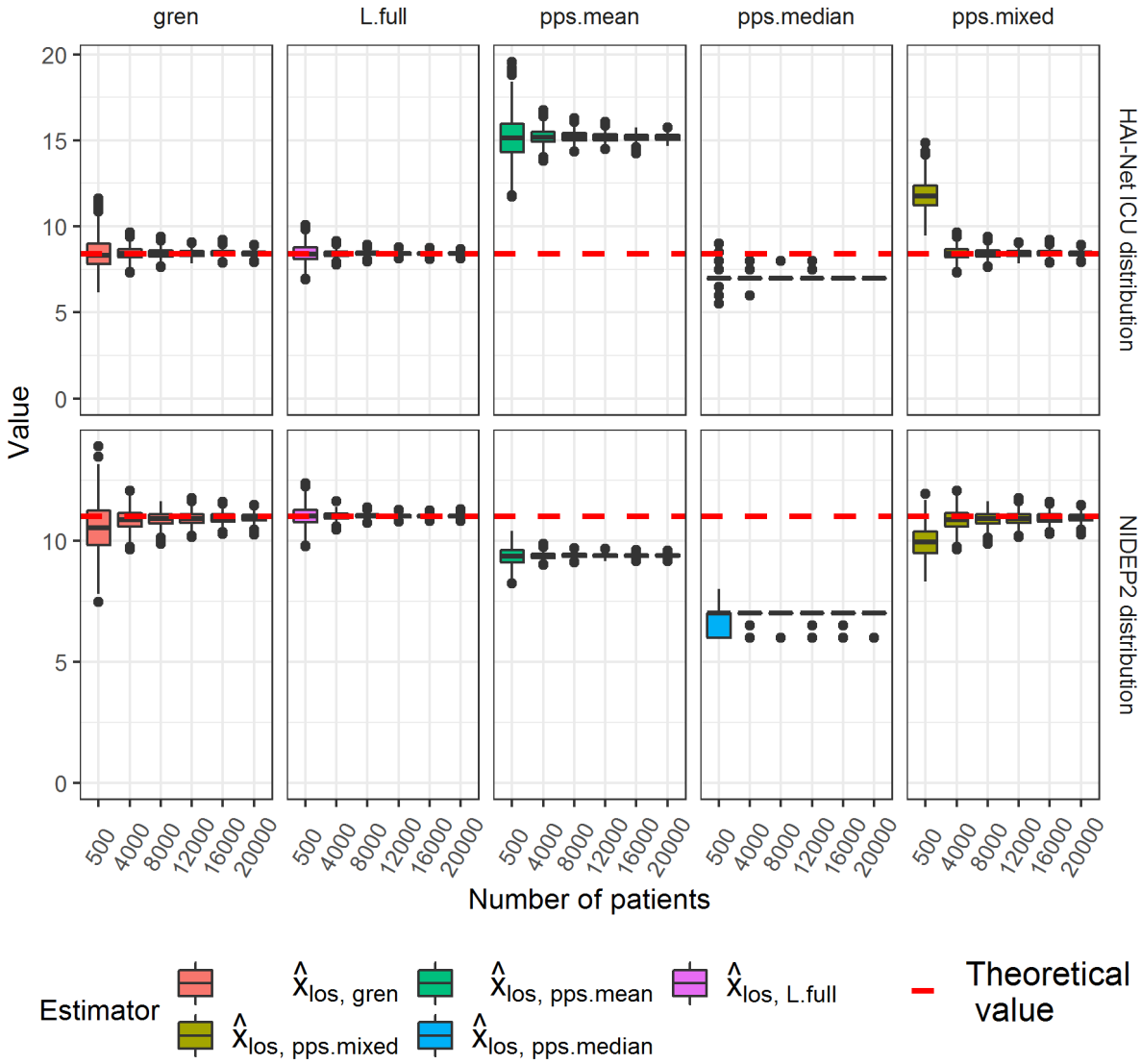

Fig. S6 *Boxplot of estimates of x_{los} for 1000 simulations each along growing number of samples of A_{los}*

References

1. Haviv M. Queues - a course in queueing theory. Springer; 2013.

2. Rhame FS, Sudderth WD. Incidence and prevalence as used in the analysis of the occurence of nosocomial infections. American Journal of Epidemiology. 1981;113:1–11.

3. Beichelt F, Fatti P. Stochastic processes and their applications. CRC Press; 2002.

4. Jankowski HK, Wellner JA. Estimation of a discrete monotone distribution. Electron J Statist. 2009;3:1567–605.

5. Gastmeier P, Bräuer H, Forster D, Dietz E, Daschner F, Rüden H. A quality management project in 8 selected hospitals to reduce nosocomial infections: A prospective, controlled study. Infection Control and Hospital Epidemiology. 2002;23:91–7.

6. European Centre for Disease Prevention and Control. European surveillance of healthcare- associated infections in intensive care units – hai-net icu protocol, version 1.02. Stockholm: ECDC; 2015.

7. "European Centre for Disease Prevention and Control". "ECDC annual epidemiological report for 2015". "ECDC Stockholm"; 2017. Available from: https://ecdc.europa.eu/sites/portal/files/documents/AER_for_2015-healthcare-associated-infections_0.pdf