University of Tartu
Faculty of Science and Technology
Institute of Mathematics and Statistics

Sten Raak

LYME DISEASE: MODELING AND ANALYZING LONG-TERM COSTS RELATED TO THE INFECTION BASED ON ESTONIAN BIOBANK DATA

Actuarial- and financial mathematics

Master's Thesis (30 ECTS)

Supervisor: PhD Erik Abner

Co-Supervisor: PhD Krista Fischer

Tartu 2025

LYME DISEASE: MODELING AND ANALYZING LONG-TERM COSTS RELATED TO THE INFECTION BASED ON ESTONIAN BIOBANK DATA

Master's thesis

Sten Raak

Abstract

Lyme disease is a prevalent vector-borne illness with significant public health implications due to its potential for multisystem effects and persistent symptoms. Understanding the associated economic burden is crucial for healthcare planning. This thesis investigates the temporal dynamics of healthcare costs surrounding a Lyme disease diagnosis, aiming to quantify whether cost increases are primarily acute or persist over a longer period, which contributes to understanding the extended healthcare services utilization potentially linked to the condition.

Using longitudinal health data from the Estonian Biobank, this study uses a relative time scale indexed to the year of first diagnosis. Linear Mixed-Effects (LME) models serve as the primary analytical framework to handle correlated repeated measures and model cost trajectories. The analysis compares diagnosed individuals to a reference group over a defined time window surrounding diagnosis. The thesis includes a background on the disease and methods, details the analysis, and presents results within the Estonian context.

CERCS research specialisation: P160 Statistics, operation research, programming, actuarial mathematics

Key Words: Mixed-effect models, repeated measurements data analysis, healthcare costs, Lyme disease, regression analysis

PUUKBORRELIOOS: NAKKUSEGA SEOTUD PIKAAJALISTE KULUDE MODELLEERIMINE JA ANALÜÜSIMINE EESTI GEENIVARAMU ANDMETE ALUSEL

Magistritöö

Sten Raak

Lühikokkuvõte

Puukborrelioos (Lyme'i tõbi) on levinud nakkushaigus, mis oma võimalike pikaajaliste tervisemõjude tõttu omab märkimisväärset tähtsust. Haigusega kaasnevate tervishoiukulude hindamine on tervishoiu planeerimisel väga tähtis. Käesolev magistritöö keskendub puukborrelioosi diagnoosimisega seotud tervishoiukulude ajalise dünaamika uurimisele. Töö peamine eesmärk on välja selgitada, kas diagnoosile järgnev kulude kasv on lühiajaline või püsib pikemalt, andes seeläbi aimu haiguse pikaajalisest mõjust tervishoiuteenuste kuludele.

Uuringus kasutati Tartu Ülikooli Eesti geenivaramu terviseandmeid ning ajaskaala keskmesse seati esimese diagnoosi aasta. Andmete analüüsimisel rakendati peamise meetodina lineaarseid segamudeleid (LME), mis sobivad hästi korduvate korreleeritud mõõtmistega terviseandmete ja kulutrajektooride modelleerimiseks. Analüüsis võrreldi diagnoosi saanud isikute ja kontrollgrupi tervishoiukulusid diagnoosieelsel ja -järgsel perioodil. Töö annab ülevaate puukborrelioosist ja LME mudelitest, kirjeldab kasutatud andmeid ja analüüsi käiku ning esitab tulemused Eesti tervishoiusüsteemi kontekstis.

CERCS teaduseriala: P160 Statistika, operatsioonianalüüs, programmeerimine, finants- ja kindlustusmatemaatika

Märksõnad: Segamudelid, kordusmõõtmiste andmete analüüs, tervishoiukulud, puukborrelioos, regressioonanalüüs

Contents

Introduction

Lyme borreliosis, commonly known as Lyme disease, is a vector-borne illness prevalent across the Northern Hemisphere. Transmitted by ticks, this disease can impact multiple organ systems, including the skin, nervous system, and joints, which can sometimes lead to persistent health problems. Its high incidence in regions like Estonia underscores its local public health importance and motivates further research into its consequences.

Beyond the clinical challenges, Lyme disease imposes a considerable economic burden on healthcare systems through costs related to diagnosis, treatment, and the management of potential long-term health effects. Understanding these costs, particularly their dynamics over time relative to when a diagnosis is made, is essential for informed healthcare planning. This thesis specifically investigates these patterns of healthcare expenditure associated with Lyme disease, focusing on how costs evolve around the time of formal diagnosis.

Such analyses often rely on longitudinal data from health registries, tracking individuals and their healthcare usage over extended periods. These datasets inherently contain repeated measurements for each person, leading to correlation among observations within individuals. This correlation structure violates the independence assumption underlying many standard statistical methods, meaning their direct application can lead to inaccurate conclusions.

To properly account for this within-subject correlation and the hierarchical nature of the data, this thesis employs Linear Mixed-Effects (LME) models. This statistical framework is well-suited for analyzing longitudinal data, as it allows for the simultaneous modeling of overall population trends (represented by fixed effects) and individual-specific deviations from those trends (represented by random effects).

The central aim of this Master's thesis is to apply LME modeling techniques to

quantify the impact of a Lyme disease diagnosis on yearly healthcare costs, using financial healthcare data from the Estonian Biobank. The analysis compares the cost trajectories of individuals diagnosed with Lyme disease against those of a relevant reference group. The investigation focuses specifically on the patterns of cost changes over a defined period spanning two years prior to and three years following the initial diagnosis event.

The subsequent chapters in this thesis will provide necessary background information on Lyme disease. Then, the thesis will present the methodology introduction, giving insight into the history and mathematical theory behind it. Following this, we will look at the results which will present the key findings from both the exploratory analysis and the final LME model.

1 Lyme disease (Lyme Borreliosis)

Lyme disease represents the most frequently reported vector-borne disease in the Northern Hemisphere (Sykes and Allen, 2011). It is recognized as an infectious disease (Shapiro, 2014), caused by spirochete bacteria belonging to the *Borrelia burgdorferi sensu lato* (s.l.) complex (Krupkaa et al., 2007). Within Europe, at least five genospecies are confirmed human pathogens: *B. afzelii*, *B. garinii*, *B. burgdorferi sensu stricto* (s.s.), *B. spielmanii*, and *B. bavariensis* (European Centre for Disease Prevention and Control (ECDC), 2016). The disease typically affects multiple organ systems, mainly the skin, nervous system, and joints (Stanek et al., 2012), although cardiac manifestations can also occur less frequently (Robinson et al., 2015). Given its potential to cause various and sometimes persistent symptoms, coupled with high incidence rates, Lyme disease constitutes a significant public health challenge that requires a thorough understanding (Embers and Narasimhan, 2013). Estonia, in particular, reports a high incidence (Figure 1), classifying the country as an endemic region for this tick-borne illness (Terviseamet, 2025).

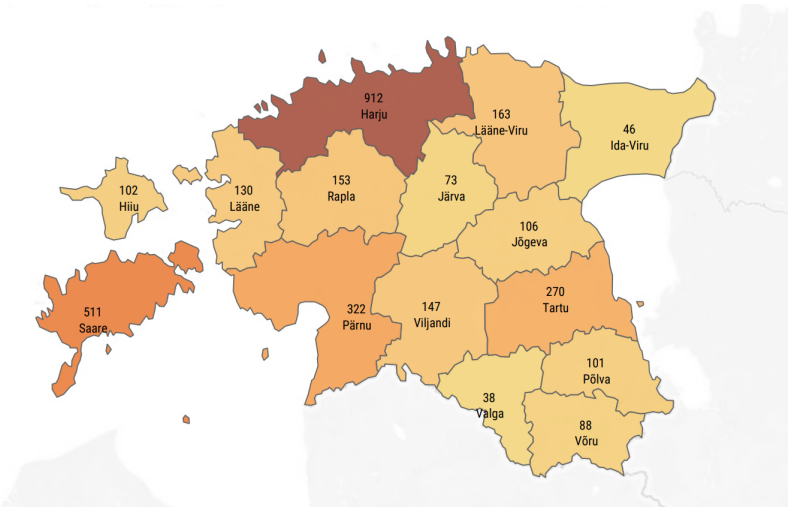

Figure 1: Registered cases of Lyme disease in Estonia by county (2024). Data Source: Estonian Health Board.

1.1 Transmission and vector ecology

Human infection with *Borrelia burgdorferi* s.l. occurs via the bite of an infected tick belonging to the genus *Ixodes* (Burgdorfer, Hayes, and Corwin, 1989). In Europe, including Estonia, the main vector species is *Ixodes ricinus*, commonly known as the castor bean tick or the sheep tick (Voyiatzaki et al., 2022). The ticks acquire the spirochetes by feeding on infected reservoir hosts during various stages of life. Small mammals, particularly rodents, along with various avian species, are considered the primary reservoir hosts responsible for maintaining the pathogen cycle in nature. (Sala, De Faveri, et al., 2016)

The lifecycle of *Ixodes* ticks typically spans 3 to 6 years, influenced by climatic conditions and host availability (Sprong et al., 2018), encompassing the stages of the egg, larval, nymph and adult (Figure 2).

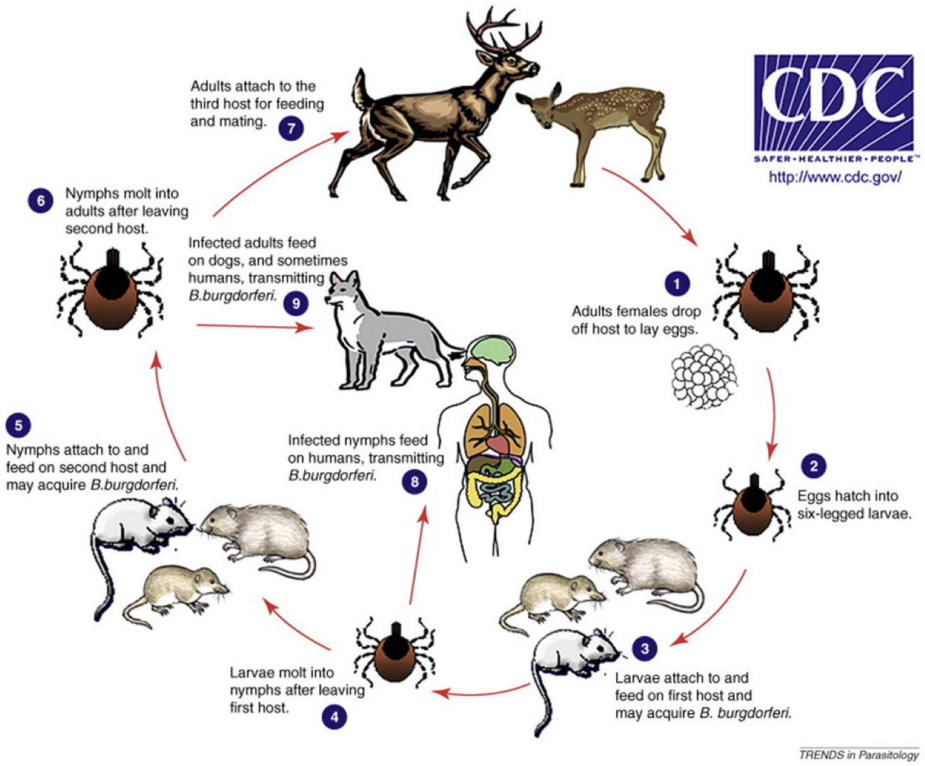

Figure 2: Lifecycle of the *Ixodes* tick. Adapted from Russell et al. (2018).

Ticks employ an ambush strategy, ascending vegetation, and waiting to attach to passing hosts. While adult ticks are capable of transmission, the nymphal stage is responsible for a large proportion of human infections, partly due to their smaller size making them less likely to be detected (Bush and Vazquez-Pertejo, 2018). Environmental factors, including changes in climate and land use patterns, can significantly influence tick population dynamics (distribution, abundance, seasonal activity), thereby altering the risk of human exposure to Lyme borreliosis. (Roome et al., 2018)

1.2 Clinical manifestations

The clinical course of Lyme disease is conventionally described in three stages: early localized, early disseminated, and late disseminated disease, although these stages can overlap and presentation varies. Each stage is typically associated with characteristic clinical features. (Bratton et al., 2008)

According to Hatchette, Davis, and Johnston (2014), early localized Lyme disease typically manifests with non-specific flu-like symptoms but is most characteristically marked by the erythema migrans (EM) skin lesion. This expanding rash (Figure 3) usually exceeds 5 cm and appears at the tick bite site within days to weeks. The diagnosis in endemic areas with typical EM is primarily clinical, as early serology is often negative.

Figure 3: Erythema Migrans rash with central clearing. Source: Centers for Disease Control and Prevention (CDC) (2024).

If the *Borrelia burgdorferi* s.l. bacteria disseminate systemically, an early disseminated disease occurs. This phase can involve multiple EM lesions, more pronounced systemic symptoms, and specific organ involvement, notably neurological complications or cardiac problems. (Schoen, 2020)

Late disseminated Lyme disease develops months to years after initial infection, if not treated, most commonly manifesting as Lyme arthritis. This typically involves intermittent or persistent inflammation in one or a few large joints, predominantly the knees. Unlike early stages, serological tests are highly sensitive for diagnosing Lyme arthritis. Less frequently, late dissemination involves the nervous system (late neuroborreliosis). (Hatchette, Davis, and Johnston, 2014)

1.3 Diagnosis

The diagnosis of Lyme disease integrates clinical presentation with laboratory findings, although clinical diagnosis can be straightforward in specific circumstances (Wright et al., 2012). Laboratory diagnostics utilize both direct methods, aimed at

detecting the *Borrelia burgdorferi* organism itself (culture detection, antigen detection, or nucleic acid amplification tests like PCR), and indirect methods, which detect the host's immune response, primarily through serological assays for specific antibodies. (Marques, 2015)

1.3.1 Clinical diagnosis

In endemic regions such as Estonia, the presentation of a typical EM rash is generally sufficient for a clinical diagnosis of early localized Lyme borreliosis, and serological testing is typically not required or recommended in this setting. Diagnosis of later stages relies more heavily on a combination of clinical signs, exposure history, and laboratory confirmation. (Prükk, Maimets, and Lutsar, 2012)

1.3.2 Laboratory testing

Laboratory tests for Lyme disease can try to find the Lyme bacteria directly (though this is less common in clinical practice due to difficulty finding the bacteria), or more typically, they look for the body's immune response to the infection by detecting specific antibodies in the blood (serology). (Marques, 2015)

Blood tests are the main laboratory tool used to diagnose later stages of Lyme disease. To improve accuracy, testing often involves two steps: first, a sensitive screening test (commonly an ELISA or similar type) is performed (Leeflang et al., 2016). If the screening test is positive or borderline, a second more specific test (usually immunoblot or Western blot) is used to confirm the result. (Parm et al., 2015)

IgM antibodies generally become detectable within 2 to 4 weeks post-infection, peak around 6-8 weeks, and may decline thereafter, though persistence can occur. IgG antibodies typically appear around 4 to 6 weeks post-infection and can remain detectable for years. However, IgM testing alone is not recommended for diagnosis beyond the first few weeks of illness due to potential for false positives and

persistence. In early neuroborreliosis, serum seropositivity rates approximate 80%. Analysis of cerebrospinal fluid for intrathecal antibody production is often more informative. It is important to recognize that prior infection does not confer lasting protective immunity, and very early antibiotic treatment can potentially abort or attenuate the antibody response. (Parm et al., 2015)

1.4 Treatment

The Lyme disease management strategy, which includes prevention and therapeutic interventions, is adapted to the specific clinical stage and manifestations (Nguyen, Cifu, and Pitrak, 2022). Optimal choices regarding antibiotic agent, dosage, route of administration, and duration of treatment remain subjects of ongoing research and refinement for certain presentations, particularly neurological and refractory forms.

For early localized Lyme disease, standard oral antibiotic regimens utilizing doxycycline, amoxicillin, or cefuroxime axetil have demonstrated high efficacy and are generally considered appropriate first-line options (Kullberg et al., 2020). However, individual patient responses can vary and there may be a possible differential susceptibility to antibiotics among various *Borrelia* genospecies may exist (Borchers et al., 2015).

Controversy surrounds the use of prolonged antibiotic courses, particularly in patients with persistent symptoms after initial treatment (see the following section) (Borchers et al., 2015). Multiple randomized controlled trials evaluating extended antibiotic therapy for such patients have generally failed to demonstrate significant additional objective benefit compared to placebo or standard treatment durations. The recommended durations of treatment typically range from 10 to 21 days for early disease, depending on the specific agent and clinical scenario. Treatment for disseminated disease typically involves longer courses and may require intravenous administration depending on severity and specific manifestation. (Schoen, 2020)

1.5 Post-Treatment Lyme Disease Syndrome

Post-Treatment Lyme Disease Syndrome (PTLDS) refers to nonspecific symptoms, including persistent fatigue, widespread musculoskeletal pain, and cognitive difficulties (often termed "brain fog"), that persist for six months or longer after the completion of standard antibiotic therapy for objectively diagnosed Lyme disease. Although sometimes used interchangeably with the less specific and often controversial term "Chronic Lyme Disease", PTLDS currently lacks universally accepted objective diagnostic criteria, hindering accurate epidemiological assessment and research. Nonetheless, given the substantial incidence of Lyme disease, the number of individuals potentially affected by PTLDS represents a significant public health concern. (Maksimyan, Syed, and Soti, 2021)

The utility of serological testing in evaluating patients with suspected PTLDS is limited and subject to significant interpretive challenges. Persistent IgM seropositivity long after initial treatment is a known phenomenon and does not reliably indicate ongoing active infection. In contrast, negative serology months or years after treated infection does not exclude PTLDS, as antibody levels can wane over time, particularly with early and effective treatment. There are currently no validated biomarkers for PTLDS. (Rebman et al., 2015)

Exactly why PTLDS develops is not fully understood. Researchers are exploring several potential reasons. These include the possibility of ongoing inflammation perhaps caused by leftover material from the bacteria, the immune system mistakenly attacking the body's own tissues after the infection, changes in how nerve pathways signal, or potentially even a persistent, hard-to-detect latent bacterial colonies. (Wong, Shapiro, and Soffer, 2022)

A prominent hypothesis involves central sensitization. This is a state where the nervous system becomes overly sensitive or hyper-reactive, possibly triggered by the initial Lyme infection and the body's inflammatory response to it. This concept is similar to hypotheses about other conditions where fatigue and pain follow

an infection. This idea is partly supported by the significant overlap in symptoms between PTLDS and conditions like fibromyalgia and chronic fatigue syndrome, which are also thought to involve central sensitization and changes in brain signaling. Currently, managing PTLDS focuses on relieving symptoms and helping patients improve their daily functioning. (Wong, Shapiro, and Soffer, 2022)

2 Linear mixed-effect models

Linear mixed-effects (LME) models provide a powerful framework for analyzing data characterized by grouping or hierarchical structures (Pinheiro and Bates, 2000). Such data structures are frequently encountered in fields like medicine, where there are repeated measurements on a single person. (Laird and Ware, 1982)

A key characteristic of such data is the grouping of data by certain factors. Examples include multiple measurements on the same individual over time, yields from experimental plots within the same block, or trait similarities among siblings from the same family. Ignoring this correlation structure and applying standard statistical models, such as Ordinary Least Squares (OLS), which assume independence of observations, can lead to inefficient parameter estimates and, critically, invalid standard errors and associated inferential statistics. (Pinheiro and Bates, 2000)

LME models explicitly account for this correlation by incorporating both fixed-effects and random-effects models. Fixed effects represent parameters associated with the entire population or specific, repeatable experimental conditions. These capture the overall average effects of interest. Random effects, conversely, are associated with individual units sampled randomly from a population. They quantify the extent to which individual units deviate from the average population behavior described by fixed effects. By incorporating random effects shared among observations within a group, LME models effectively represent the covariance structure induced by the data hierarchy. (Pinheiro and Bates, 2000)

While fixed-effect models can capture differences between observed groups, they limit generalizability. Specifically, inferences are restricted to the particular groups included in the study and do not allow estimation of the variability between groups within the broader population from which the sample was drawn. LME models overcome these limitations by simultaneously modeling both population-average trends and the structure of variability, encompassing variation both within and between groups. (Laird and Ware, 1982)

2.1 Historical context and foundations

The concepts of mixed-effects models and variance component analysis trace back to early work in quantitative genetics. A foundational analysis of the correlation between relatives, based on the principles of Mendelian inheritance, was provided by Fisher (1918). In this work, Fisher introduced the concept of partitioning the total phenotypic variance of a trait into distinct components attributable to different underlying causes.

He emphasized the importance of analyzing the variance itself and proposed a decomposition of this variance into contributions primarily from additive genetic effects and dominance deviations, while acknowledging potential roles for environment. Furthermore, Fisher demonstrated that under Mendelian inheritance, the expected correlations between different types of relatives are dependent upon the relative magnitudes of these identified variance components. Specifically, he showed that the influence of dominance effects tends to reduce the parental correlation to a greater extent than the fraternal correlation, thereby offering a potential statistical method for distinguishing the effects of genetic dominance from those of shared environmental influences. This work by Fisher established the fundamental principle of analyzing variance components as a means to understand the different sources contributing to variation within populations.

Subsequent developments in statistical methodology, driven in part by the challenges encountered when applying traditional multivariate models to longitudinal or repeated measures data, which often suffer from imbalance or high dimensionality, led to the more formal development and popularization of random effects models, notably by Laird and Ware (1982). These models were designed to explicitly address distinct sources of variation, namely variability within-subject and between-subject. This provides a robust framework suitable for analyzing complex serial measurement data.

2.2 Model formulation

A general and widely applicable framework for LME models, particularly suited for longitudinal or repeated measures data, is the two-stage random-effects model formulation described by Laird and Ware (1982) and Harville (1977), on which the next three sections are based on.

2.2.1 The Two-Stage model

This formulation conceptually separates the modeling process into within-individual and between-individual components:

Stage 1: Individual level model. For each individual unit i (where $i = 1, \ldots, m$), the $n_i \times 1$ vector of responses $\boldsymbol{y}_i$ is modeled as:

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\alpha} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{e}_i \tag{1}$$

where:

- $\boldsymbol{\alpha}$ is a $p \times 1$ vector of unknown fixed-effect parameters, common to all individuals.
- $\boldsymbol{X}_i$ is a known $n_i \times p$ design matrix relating $\boldsymbol{\alpha}$ to $\boldsymbol{y}_i$.
- $\boldsymbol{b}_i$ is a $k \times 1$ vector of unknown random effects, specific to individual i.
- $\boldsymbol{Z}_i$ is a known $n_i \times k$ design matrix relating $\boldsymbol{b}_i$ to $\boldsymbol{y}_i$.
- $\boldsymbol{e}_i$ is an $n_i \times 1$ vector of random errors for individual i.

Conditional on $\boldsymbol{\alpha}$ and $\boldsymbol{b}_i$, the errors $\boldsymbol{e}_i$ are assumed to follow a multivariate normal distribution, $\boldsymbol{e}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{R}_i)$, where $\boldsymbol{R}_i$ is an $n_i \times n_i$ positive-definite covariance matrix. The errors $\boldsymbol{e}_i$ are typically assumed independent across individuals. Often,

a simpler structure $\boldsymbol{R}_i = \sigma^2 \boldsymbol{I}_{n_i}$ is assumed, implying conditional independence and homoscedasticity of measurements within an individual, given their random effects.

Stage 2: Population level model. The individual random effects $\boldsymbol{b}_i$ are assumed to vary across individuals according to a multivariate normal distribution:

$$\boldsymbol{b}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{D}) \tag{2}$$

where $\boldsymbol{D}$ is a $k \times k$ positive-definite covariance matrix, representing the variance and covariance of the random effects in the population. The random effects $\boldsymbol{b}_i$ are assumed independent across individuals and independent of the error vectors $\boldsymbol{e}_i$. The fixed effects $\boldsymbol{\alpha}$ are treated as fixed, unknown constants.

2.2.2 Marginal distribution and covariance structure

By integrating out the random effects $\boldsymbol{b}_i$, the marginal distribution of the response vector $\boldsymbol{y}_i$ for individual i is also multivariate normal:

$$\boldsymbol{y}_i \sim \mathcal{N}(\boldsymbol{X}_i\boldsymbol{\alpha}, \boldsymbol{V}_i) \tag{3}$$

where the marginal covariance matrix $\boldsymbol{V}_i$ is given by:

$$\boldsymbol{V}_i = \mathrm{Var}(\boldsymbol{y}_i) = \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i^T + \boldsymbol{R}_i \tag{4}$$

This marginal covariance matrix $\boldsymbol{V}_i$ is crucial as it captures both the within-individual variability (through $\boldsymbol{R}_i$) and the between-individual variability (through the term involving $\boldsymbol{D}$), explicitly modeling the dependence structure among observations within the same individual. The primary parameters to be estimated in the model are the fixed effects $\boldsymbol{\alpha}$ and the variance components defining the structure and elements of $\boldsymbol{R}_i$ and $\boldsymbol{D}$ (collectively denoted as $\boldsymbol{\theta}$).

2.2.3 Special cases and examples

This general formulation encompasses many commonly used models:

- **Random intercept model:** If only the intercept term varies randomly across individuals, then $\boldsymbol{b}_i = b_{0i}$ is scalar ($k = 1$), $\boldsymbol{Z}_i$ is an $n_i \times 1$ vector of ones, and $\boldsymbol{D} = \sigma_b^2$ is a scalar variance. If additionally $\boldsymbol{R}_i = \sigma^2 \boldsymbol{I}$, the model for observation j on individual i becomes $y_{ij} = (\boldsymbol{X}_i \boldsymbol{\alpha})_j + b_{0i} + e_{ij}$. In this case, $\mathrm{Var}(y_{ij}) = \sigma_b^2 + \sigma^2$ and the covariance between any two distinct measurements on the same individual is $\mathrm{Cov}(y_{ij}, y_{ij'}) = \sigma_b^2$ for $j \neq j'$. This corresponds to a compound symmetry covariance structure, often referred to as a simple random-intercept model.

- **Random intercept and slope model:** This is commonly applied to growth curve data or longitudinal studies where y_{ij} is the measurement at time t_{ij}. Here, $\boldsymbol{b}_i = (b_{0i}, b_{1i})^T$ might represent subject-specific deviations in intercept and slope from the population average trajectory ($k = 2$). The design matrix $\boldsymbol{Z}_i$ would typically have columns $\mathbf{1}$ and $(t_{i1}, \ldots, t_{in_i})^T$. The covariance matrix $\boldsymbol{D}$ would be a 2×2 matrix containing the variances $\mathrm{Var}(b_{0i})$ and $\mathrm{Var}(b_{1i})$, and the covariance $\mathrm{Cov}(b_{0i}, b_{1i})$.

- **Nested Models:** Models for experimental designs like split-plots, or multi-level data with more than two levels (for example students within classrooms within schools), involve defining nested random effects. This fits within the general framework by appropriately structuring the random effects vector $\boldsymbol{b}_i$ and its design matrix $\boldsymbol{Z}_i$ to represent multiple hierarchical levels of random variation.

2.3 Estimation approaches

Estimating the model parameters (fixed effects $\boldsymbol{\alpha}$ and variance components $\boldsymbol{\theta}$) typically involves likelihood-based methods applied to the marginal distribution of $\boldsymbol{y} = (\boldsymbol{y}_1^T, \ldots, \boldsymbol{y}_m^T)^T$.

2.3.1 Maximum Likelihood (ML)

ML estimation identifies parameter values $(\hat{\boldsymbol{\alpha}}_{ML}, \hat{\boldsymbol{\theta}}_{ML})$ that maximize the likelihood function derived from the marginal multivariate normal distributions of the $\boldsymbol{y}_i$. For a fixed value of the variance components $\boldsymbol{\theta}$, the ML estimate of $\boldsymbol{\alpha}$ corresponds to the generalized least squares (GLS) estimator:

$$\hat{\boldsymbol{\alpha}}(\boldsymbol{\theta}) = \left(\sum_{i=1}^{m} \boldsymbol{X}_i^T \boldsymbol{V}_i^{-1}(\boldsymbol{\theta}) \boldsymbol{X}_i \right)^{-1} \left(\sum_{i=1}^{m} \boldsymbol{X}_i^T \boldsymbol{V}_i^{-1}(\boldsymbol{\theta}) \boldsymbol{y}_i \right) \tag{5}$$

where $\boldsymbol{V}_i(\boldsymbol{\theta}) = \boldsymbol{R}_i(\boldsymbol{\theta}) + \boldsymbol{Z}_i \boldsymbol{D}(\boldsymbol{\theta}) \boldsymbol{Z}_i^T$. Since $\boldsymbol{\theta}$ (which determines $\boldsymbol{V}_i$) is generally unknown, $\boldsymbol{\alpha}$ and $\boldsymbol{\theta}$ must typically be estimated simultaneously or iteratively using numerical optimization algorithms. A known issue with ML estimation is that the resulting estimates of variance components $(\hat{\boldsymbol{\theta}}_{ML})$ tend to be biased downwards, particularly in studies with a small number of groups (m), because the estimation process does not fully account for the degrees of freedom lost in estimating the fixed effects $\boldsymbol{\alpha}$. (Laird and Ware, 1982)

2.3.2 Restricted Maximum Likelihood (REML)

While standard Maximum Likelihood (ML) estimation can be used for LME models, its estimates for the variance components are known to be biased, often underestimating the true variance, especially when the number of groups or subjects is relatively small. Restricted Maximum Likelihood (REML) estimation was specifically developed to address this issue and provide less biased estimates for these

crucial variance components (Laird and Ware, 1982).

REML achieves this improved estimation by modifying the likelihood function used in the optimization process. Conceptually, REML focuses purely on the variance components by mathematically adjusting the likelihood calculation to remove the influence of simultaneously estimating the fixed effects. An alternative way to understand REML arises from a Bayesian viewpoint: it can be seen as maximizing the likelihood of the variance components $\boldsymbol{\theta}$ after effectively averaging over all possible values of the fixed effects $\boldsymbol{\alpha}$, typically assuming we have no strong prior information about $\boldsymbol{\alpha}$. (Laird and Ware, 1982; Harville, 1977)

Because REML generally produces more accurate and less biased estimates of variance components compared to ML, it has become the standard and preferred method for fitting LME models, particularly when the variances themselves are of interest or when making inferences that depend heavily on them (Laird and Ware, 1982). The estimation procedure typically involves two conceptual steps:

1. REML estimation is used first to find the optimal estimate for the variance parameters, denoted $\hat{\boldsymbol{\theta}}_{REML}$.

2. Then, this REML variance estimate ($\hat{\boldsymbol{\theta}}_{REML}$) is treated as if it were the true value, and it is substituted back into the standard formula for estimating the fixed effects, $\boldsymbol{\alpha}$, such as the Generalized Least Squares (GLS) equation presented earlier (5) (Pinheiro and Bates, 2000).

3 The bootstrap method

The bootstrap method, as introduced and comprehensively described by Efron (1979), on which the entire bootstrap section is based, addresses a fundamental challenge in statistical inference. The method provides a means to estimate the sampling distribution of a statistic, or related quantities derived from it, when the underlying probability distribution F generating the data is unknown and only a single data sample $x = (X_1, X_2, \ldots, X_n)$ drawn from F is available. Specifically, interest often lies in characterizing the distribution of a statistic $R(x, F)$, which may depend on both the sample x and the unknown distribution F.

Efron (1979) presented the bootstrap as a general, non-parametric methodology for this purpose, suggesting its potential for broader applicability and robustness compared to preceding resampling techniques like the jackknife. For instance, the bootstrap provides consistent estimates for the variance of statistics such as the sample median, a known limitation of the standard jackknife procedure. Furthermore, the jackknife can be mathematically demonstrated to be a linear approximation to the bootstrap.

3.1 Core bootstrap workflow

For the common scenario involving a single observed sample, the bootstrap technique typically proceeds through the following stages:

1. **Construct the Empirical Distribution Function ($\hat{F}$):** Based on the observed sample $x = (x_1, x_2, \ldots, x_n)$, define the empirical distribution function $\hat{F}$. This discrete distribution assigns probability mass $1/n$ to each observed data point x_i. $\hat{F}$ serves as a non-parametric estimate of the true, unknown distribution F.

2. **Generate Bootstrap Samples:** Draw n observations with replacement

from the original data set $\{x_1, x_2, \ldots, x_n\}$ to form a bootstrap sample, denoted $x^* = (x_1^*, x_2^*, \ldots, x_n^*)$. Sampling with replacement is important, it implies that original data points may appear multiple times, or not at all, in any given bootstrap sample. This resampling process is equivalent to drawing a random sample of size n from the empirical distribution $\hat{F}$.

3. **Calculate the Statistic for the Bootstrap Sample:** Compute the statistic of interest, R, using the generated bootstrap sample x^*. If R depends explicitly on the distribution, use $\hat{F}$. Denote this calculated value as $R^* = R(x^*, \hat{F})$.

4. **Repeat and Approximate the Sampling Distribution:** Repeat steps 2 and 3 a large number of times, B, thereby generating B independent bootstrap samples and calculating the corresponding statistic R^{*b} for each sample ($b = 1, \ldots, B$). The empirical distribution of these B replicate values $\{R^{*1}, R^{*2}, \ldots, R^{*B}\}$, often referred to as the bootstrap distribution, serves as an approximation to the true sampling distribution of the original statistic $R(x, F)$.

3.2 Justification based on the plug-in principle

The theoretical justification for using the distribution of bootstrap replicates $\{\hat{\theta}^{*b}\}$ to approximate the true sampling distribution of $\hat{\theta}$ is grounded in the plug-in principle. The idea is to estimate properties of the true, unknown distribution F, such as the mean or standard error of $\hat{\theta}$, by replacing F with the empirical distribution $\hat{F}$ formed from the observed sample x. The bootstrap implements this by drawing repeated samples from $\hat{F}$ and recomputing $\hat{\theta}$, thereby approximating the variability in $\hat{\theta}$ as if we were sampling from the true distribution.

The standard error of the statistic $\hat{\theta} = S(x)$ is a functional of the true distribution F, denoted $SE_F(\hat{\theta})$. The bootstrap estimates this quantity by applying the same

functional definition but substituting $\hat{F}$ for F. This corresponds to calculating the standard deviation of the B bootstrap replicates $\{\hat{\theta}^{*b}\}$. The bootstrap estimate of the standard error is thus given by:

$$\hat{SE}_{boot} = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}(\hat{\theta}^{*b} - \bar{\theta}^{*})^2} \tag{6}$$

where $\bar{\theta}^{*} = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}^{*b}$ represents the mean of the B bootstrap replicates.

A key theoretical result about the bootstrap is that, under suitable regularity conditions on the statistic S and the distribution F, the distribution of the centered bootstrap replicates $(\hat{\theta}^{*} - \hat{\theta})$ converges (as $n \to \infty$) to the same limiting distribution as the centered true sampling error $(\hat{\theta} - \theta)$. This convergence implies that the bootstrap distribution effectively mimics the characteristics (such as shape, spread, and bias) of the true sampling distribution. Consequently, $\hat{SE}_{boot}$ serves as a consistent estimator for the true standard error $SE_F(\hat{\theta})$ under fairly general conditions.

Essential assumptions for this theoretical justification include:

- The original sample x is representative of the underlying population F.
- The sample size n is sufficiently large to ensure $\hat{F}$ is a good approximation of F and for asymptotic results to hold.
- The statistic $\hat{\theta}$ is sufficiently "smooth" as a function of the data.

4 Data description

The data utilized for this thesis were sourced from the Estonian Biobank. At the time of the current data freeze, the Estonian Biobank had 212 955 participants. All individuals who joined the biobank provided broad informed consent, permitting the use of their health data for a wide range of research purposes.

Information on diagnoses, coded according to the International Classification of Diseases, is obtained through regular and systematic linking of biobank participants' data with national health registries, primarily the Estonian Health Insurance Fund, as well as other relevant national health databases. The majority of the electronic health records utilized have been collected since 2004, providing a significant amount of data for analysis.

The healthcare costs analyzed in this thesis encompass three categories: Primary care costs, outpatient care costs and inpatient care costs. It is important to note that these selected categories, while foundational, represent approximately 36% of the total healthcare expenditures recorded within the Estonian Biobank data used. Outpatient medications, highly specialized laboratory tests billed separately, or distinct surgical procedure costs are not included in this thesis as the detailed categorization of such diverse elements was beyond the defined scope of the current project. However, these three primary cost types are considered to provide a substantially valid representation of the core healthcare engagement.

For the specific analyzes conducted in this thesis, the occurrence of diagnostic events for Lyme disease and selected comparator conditions was ascertained using the following ICD-10-based categorizations from the Estonian Biobank data:

- Lyme disease: ICD-10 code A69.2
- Impetigo: ICD-10 code L01
- Follicular cysts of skin and subcutaneous tissue: ICD-10 code L72

- Malignant neoplasm of breast (breast cancer): ICD-10 code C50
- Type 2 diabetes: ICD-10 code E11

5 Exploratory data analysis

Exploratory data analysis (EDA) is a crucial preliminary phase preceding formal statistical modeling. Its primary objective is to gain familiarity with the data by examining its structure, identifying potential patterns or trends, detecting anomalies or outliers.

A primary focus of this analysis is to understand the evolution of healthcare costs associated with Lyme disease diagnosis. For this, the trend of yearly healthcare costs were examined, stratified by diagnosis status, as illustrated in Figure 4.

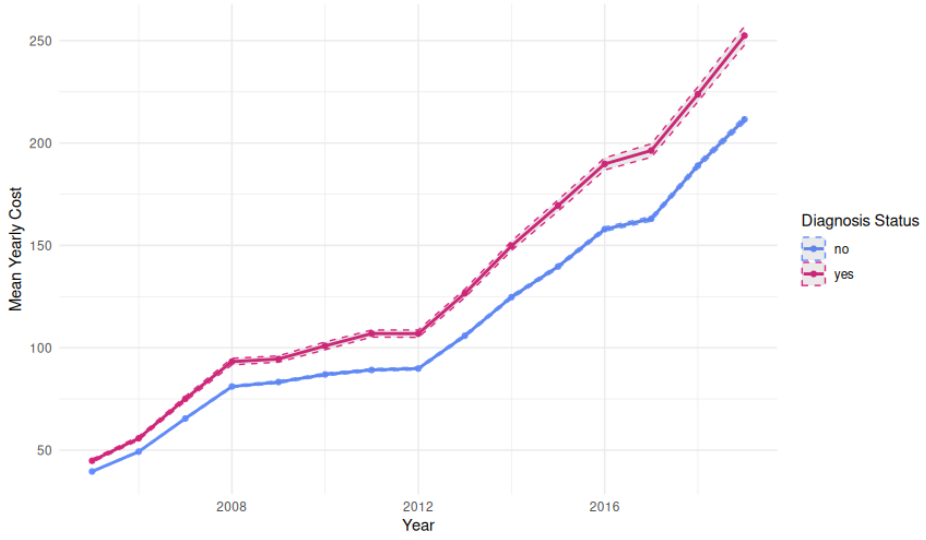

Figure 4: Trend of yearly healthcare costs stratified by Lyme diagnosis status (in euros).

Figure 4 plots the average yearly healthcare costs against the calendar year. A distinct upward trend in yearly costs is observable for both diagnosed and non-diagnosed groups across the study period. In particular, after the initial years, the mean costs for individuals diagnosed with Lyme disease appear to be consistently elevated compared to those without the diagnosis. This visual evidence suggests a significant difference in costs, motivating the use of appropriate statistical models, such as Linear Mixed-Effects models, to formally quantify these differences.

Furthermore, the potential influence of diagnosis frequency or cumulative burden on cost patterns warrants investigation. A greater number of recorded diagnoses could signify more severe, persistent, or recurrent disease states, plausibly leading to increased healthcare utilization and associated costs. Figure 5 explores this by presenting the yearly healthcare costs stratified by the total count of Lyme diagnoses per individual.

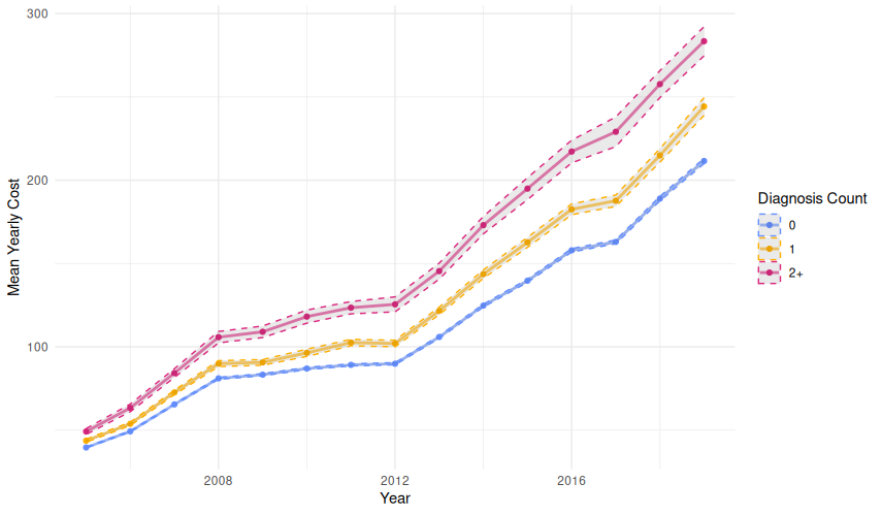

Figure 5: Trend of yearly mean primary care costs stratified by count of Lyme diagnoses (in euros).

Similarly to the preceding figure, Figure 5 displays the average annual costs over time, stratified by the cumulative count of Lyme diagnoses. We only count a diagnosis as an additional if 12 months have passed since the last one. Individuals with zero diagnoses have the lowest average costs, and costs generally demonstrate a monotonic increase with the number of diagnoses. All groups show an increasing trend over calendar time. The width of the confidence intervals is mostly reflected by the sample size limitation within each stratum, being wider for groups with fewer individuals.

Analyzing cost trends by calendar year provides population-level insights but confounds individuals at different stages relative to their disease onset. To isolate

the cost dynamics specifically surrounding the diagnosis event, an alternative perspective aligns individuals based on the year of their first recorded Lyme disease diagnosis. Time is therefore redefined as years relative to this diagnosis event (Year 0), facilitating the examination of cost patterns preceding and succeeding the diagnosis. Figure 6 visualizes average total costs using this relative time scale.

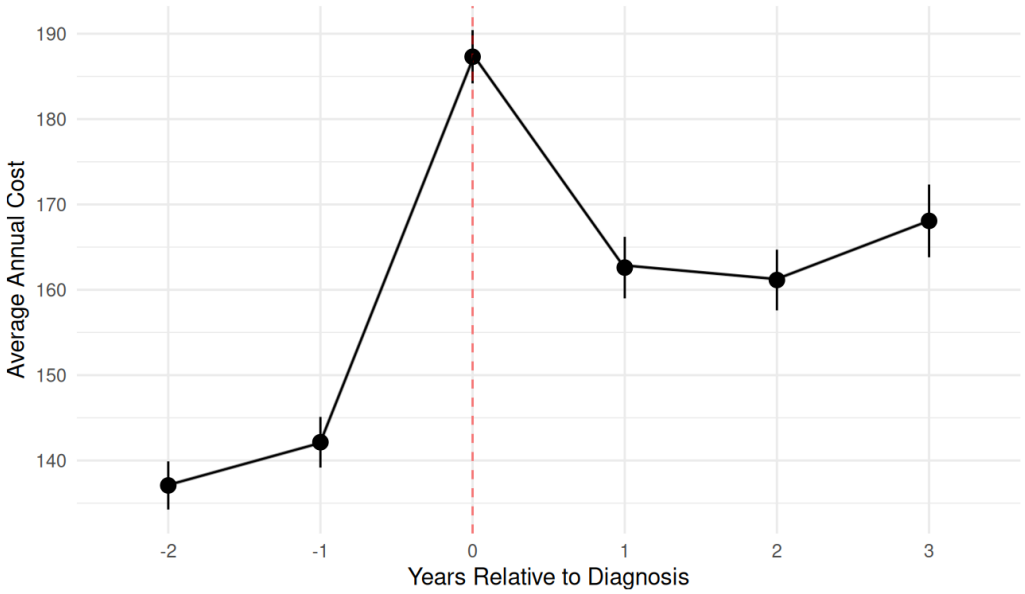

Figure 6: Average total costs relative to Lyme disease diagnosis (in euros).

Figure 6 presents the average annual total costs plotted against time relative to the initial Lyme disease diagnosis year. The time axis spans from two years prior to five years post-diagnosis. Examination of the trends reveals a significant increase in costs coinciding with the diagnosis year. Subsequent to diagnosis, average costs appear to remain persistently elevated compared to prediagnosis levels for several years.

6 Modeling healthcare costs

The exploratory analysis, particularly the visualization of costs relative to the diagnosis year (Figure 6), highlighted distinct cost patterns and potential differences between cohorts. To quantify these observations and test hypotheses regarding the impact of diagnosis over time, statistical modeling is used.

The longitudinal structure of the data, characterized by repeated cost measurements for each individual across multiple years, renders Linear Mixed-Effects (LME) models particularly suitable. LME models accommodate the inherent correlation among repeated observations within individuals while enabling the estimation of population-average effects (fixed effects) associated with covariates such as time relative to diagnosis and diagnosis status.

6.1 Defining the analytical time window

A preliminary step involved selecting the optimal time window relative to the diagnosis year for inclusion in the models. The goal was to capture relevant pre- and post-diagnosis dynamics while maintaining sufficient statistical power. A pre-diagnosis period including Years -2 and -1 was chosen to provide a baseline. To determine the post-diagnosis duration, models considering follow-up periods from +2 to +7 years were evaluated based on the resulting sample size (Table 1).

Table 1: Sample Size (N Individuals) by Pre- and Post-Diagnosis Follow-up Window.

Relative Year Range	**Sample Size (N)**
$[-2, +2]$	10 995
$[-2, +3]$	9 880
$[-2, +4]$	9 011
$[-2, +5]$	8 117
$[-2, +6]$	7 142
$[-2, +7]$	6 330

Given the complexity of LME models, particularly those including interaction terms, a substantial sample size is advantageous for model stability and reliable estimation. As Table 1 shows, extending the follow-up beyond Year +3 led to a significant decrease in sample size (to around 9 000). Therefore, the range spanning Year -2 to Year +3 was selected for subsequent analyses. This window retains a large sample (N = 9 880) while providing a sufficiently informative post-diagnosis observation period.

6.2 Baseline model for diagnosed individuals

As an initial step in the modeling process, a Linear Mixed-Effects (LME) model was fitted using data exclusively from individuals diagnosed with Lyme disease. The objective of this baseline model was to characterize the average trajectory of yearly healthcare costs within this specific group, while accounting for key demographic factors (sex, age at diagnosis) and inherent individual-level variability, prior to making comparisons with a non-diagnosed cohort.

The model specified mean yearly healthcare costs as the outcome, predicted by fixed effects for relative year (categorical, with Year -2 as the reference level), biological sex, and age at diagnosis. A random intercept was included for each individual to capture subject-specific baseline cost levels. The model parameters were estimated using Restricted Maximum Likelihood (Table 2).

The interpretation of the intercept should be made cautiously due to the centering of covariates (age 0 is not meaningful), but it serves as the statistical baseline. The results clearly indicate that, within the diagnosed group, average yearly healthcare costs increase significantly across all subsequent years compared to the baseline level two years prior to diagnosis (Year -2). The most substantial increases relative to baseline occur in the year of diagnosis (Year 0) and three years post-diagnosis (Year 3). Demographic factors also show significant associations. Being female is associated with significantly higher average annual costs compared to males within

Table 2: Fixed Effects Estimates: Diagnosed Group Baseline Model (estimates in euros)

Predictor	**Estimate**	**Std. Error**	**p-value**
Intercept	-21.798	4.144	<0.001
Year -1	15.028	1.999	<0.001
Year 0	68.720	1.999	<0.001
Year 1	53.167	1.999	<0.001
Year 2	60.570	1.999	<0.001
Year 3	77.110	1.999	<0.001
Sex (Female vs Male Ref.)	24.137	2.589	<0.001
Age at Diagnosis	2.208	0.073	<0.001

this cohort, and each additional year of age at the time of diagnosis corresponds to a statistically significant increase in estimated costs.

Furthermore, the random effects estimates highlight substantial heterogeneity among individuals. The standard deviation associated with the random intercepts (99.23) is considerable relative to the fixed effect estimates, indicating large differences in baseline healthcare spending levels between diagnosed individuals. The residual standard deviation (140.47) signifies the marked variability in costs within the same individual over time, even after accounting for the model's predictors and individual baseline differences.

While this model provides valuable insights into cost patterns and variability solely within the Lyme-diagnosed population, it does not, by design, permit direct comparisons to individuals without the diagnosis. This comparison is addressed in subsequent models.

6.3 Direct comparison using group interaction model

To directly compare the cost trajectories between individuals diagnosed with Lyme disease and non-diagnosed individuals, a second LME model was employed. This

analysis utilized a combined dataset where non-diagnosed individuals were assigned a random index year to align them on the same relative time scale (Year -2 to Year +3) as the diagnosed group. This enables a fair and simple comparison between the groups.

The core of this model was the inclusion of an interaction term between the categorical relative year factor and the diagnosis group indicator (no diagnosis vs Lyme disease diagnosis). This interaction allows for testing whether the temporal pattern of costs significantly differs between the two groups. As with the previous model, a random intercept for each individual was included to account for subject-level variability.

The fixed effects estimates for this interaction model are presented in Table 3 below.

Table 3: Fixed effects estimates: group comparison model with interaction (estimates in euros)

Predictor	**Estimate**	**Std. Error**	**p-value**
Intercept	81.465	0.520	<0.001
Year -1	9.999	0.600	<0.001
Year 0	20.978	0.600	<0.001
Year 1	29.809	0.600	<0.001
Year 2	42.402	0.600	<0.001
Year 3	54.538	0.600	<0.001
Group Lyme	16.779	1.865	<0.001
Year -1 * Lyme	5.031	2.153	0.019
Year 0 * Lyme	47.682	2.153	<0.001
Year 1 * Lyme	23.294	2.153	<0.001
Year 2 * Lyme	18.170	2.153	<0.001
Year 3 * Lyme	22.558	2.153	<0.001

The model intercept estimates the average yearly cost for the non-diagnosed reference group at the baseline time point (Year -2). The main effects for relative year indicate that costs significantly increase over time for this reference group

compared to their baseline. The main effect for the Lyme diagnosis group reveals that diagnosed individuals generally have increased costs at all times, even at the baseline.

The interaction terms are essential for understanding how the difference between the groups evolves. They represent the additional change in cost for the Lyme group in a given year, compared to the change experienced by the reference group in that same year. The results show a significant interaction in Year 0, indicating a substantial additional cost spike specifically associated with the diagnosis year for the Lyme group, over and above their baseline difference and the general time trend. Significant positive interactions persist in Years 1, 2, and 3, demonstrating that the cost gap between the Lyme group and the reference group remains significantly larger during the post-diagnosis period compared to the difference observed at baseline.

This model directly addresses the research question by testing and quantifying the interaction between diagnosis status and relative time. It confirms that the impact of Lyme disease on costs is not simply a constant baseline shift but involves a distinct pattern, especially pronounced around the diagnosis year, with lasting effects. While calculating the total estimated difference at any specific time point requires combining the main group effect and the relevant interaction term, the significance of the interactions clearly demonstrates differing cost trajectories.

7 Alternative two-stage approach: modeling excess costs

An alternative strategy was explored to directly model the excess cost potentially attributable to Lyme disease. This involved a two-stage approach:

1. A model predicting expected costs was fitted using data only from the non-diagnosed group (Appendix 1).
2. This model was used to predict the expected yearly costs for individuals in the diagnosed group, representing their hypothetical costs had they followed the pattern of the non-diagnosed group.
3. The difference between the observed yearly costs and these predicted costs for the diagnosed individuals was calculated. This difference aims to represent the excess cost associated with Lyme disease status.
4. The final LME model was then fitted using this calculated difference as the outcome variable for the diagnosed group.

7.1 Simple excess cost model

The initial model within this two-stage framework examined the temporal pattern of the calculated excess cost. A simple LME model was specified, predicted only by the categorical relative year factor with Year -2 as the reference level. A random intercept for each individual was included to account for person-specific average levels of excess cost.

The results (Table 4) suggest that, on average, there was a small but statistically significant positive excess cost even two years prior to the index Lyme disease diagnosis (represented by the intercept). The change in excess cost in the year immediately before diagnosis was not significantly different from the baseline level.

Table 4: Fixed effects estimates: simple LME model of excess cost (estimates in euros).

Predictor	Estimate	Std. Error	p-value
Intercept	4.151	1.728	0.016
Year -1	3.629	1.994	0.069
Year 0	45.192	1.994	<0.001
Year 1	19.077	1.994	<0.001
Year 2	14.743	1.994	<0.001
Year 3	15.968	1.994	<0.001

However, a substantial and highly significant increase in excess cost occurred in the year of diagnosis. Significant positive excess costs relative to the baseline persisted throughout the post-diagnosis period. Although these were notably lower than the peak observed in Year 0.

This model directly estimates the average trajectory of the calculated excess cost over time. However, its primary limitation is the lack of adjustment for demographic or clinical covariates that might also influence the magnitude of this cost difference between observed and expected values. This limitation is addressed in the subsequent model refinement.

7.2 Refined excess cost model with covariates

To gain a more nuanced understanding of the excess costs associated with Lyme disease, the analysis was refined by fitting a more comprehensive LME model to the calculated cost difference. This model incorporated demographic covariates, biological sex and age at diagnosis (centered at 40 years), and focused specifically on the year of diagnosis and the three subsequent years. By omitting a general intercept term from the model, the coefficients for the relative year variables directly estimate the average excess cost for each specific year, conditional on the included covariates. Fitting was performed using Restricted Maximum Likelihood (REML),

and individual-level variation was captured with a random intercept.

Table 5 presents the fixed effects estimates from this adjusted model. The coefficients for the relative years now directly estimate the average excess cost during that specific year for a reference individual (defined here as a male diagnosed at age 40).

Table 5: Fixed effects estimates: adjusted excess cost model (estimates in euros)

Predictor	**Estimate**	**Std. Error**	**p-value**
Year 0	48.635	2.758	<0.001
Year 1	22.602	2.758	<0.001
Year 2	19.985	2.758	<0.001
Year 3	22.827	2.758	<0.001
Sex (Female vs Male Ref.)	-7.251	2.827	0.010
Age (centered at 40)	0.249	0.080	0.002

The results highlight a clear pattern in the adjusted excess costs. The financial impact peaks significantly in the year of diagnosis, where the average excess cost is estimated to be around 49 units for the reference individual. While this impact decreases in the following year, the excess costs remain statistically significant and substantial throughout the subsequent three years studied. This finding strongly confirms the persistence of an economic burden associated with Lyme disease, even after adjusting for key demographic factors.

This adjusted model also sheds light on how demographic factors relate to the magnitude of these excess costs. Compared to the male reference group, females were estimated to have now a lower average excess cost. Furthermore, older age at diagnosis showed a positive association, with each year above age 40 corresponding to a small but statistically significant increase in the estimated excess costs.

There remains considerable differences between individuals in their overall level of

excess healthcare costs, as well as notable fluctuations within individuals over time that are not captured by the current predictors. This highlights the inherent heterogeneity in how individuals experience and utilize healthcare resources following a Lyme disease diagnosis.

7.3 Bootstrapping the final model

To further assess the stability of the parameter estimates obtained from the refined excess cost model above and to derive empirical confidence intervals, a non-parametric bootstrap procedure was implemented. The process involved resampling individuals with replacement from the original dataset and refitting the LME model to each resampled dataset. This procedure was repeated B=100 times. While a larger number of bootstrap replicates is typically recommended for robust inference, these initial results provide a preliminary assessment of estimate stability.

The mean coefficient estimates, empirical standard errors (SE), and percentile-based 95% confidence intervals (lower limit - LL, upper limit - UL) across the 100 bootstrap replicates are summarized in Table 6.

Table 6: Bootstrap results (B=100 replicates): fixed effects for refined excess cost model (estimates in euros)

Predictor	**Mean**	**SE**	**95% CI (LL)**	**95% CI (UL)**
Year 0	48.867	2.233	44.491	53.243
Year 1	22.801	2.380	18.136	27.466
Year 2	20.081	2.209	15.571	24.411
Year 3	22.647	2.884	16.995	28.300
Sex	-7.281	2.327	-11.842	-2.719
Age (centered at 40)	0.242	0.074	0.098	0.386

Comparing the bootstrap results (Table 6) with the original estimates (Table 5), we observe general consistency. The bootstrap mean estimates are reasonably close

to the point estimates obtained from the original model fit. The bootstrap standard errors are also broadly comparable to the values calculated by the final model.

The 95% bootstrap confidence intervals support the conclusions drawn from the original model regarding statistical significance. The intervals for all relative year coefficients are entirely above zero, reinforcing the finding of significant higher excess costs in the diagnosis year and the three subsequent years. Similarly, the confidence interval for the effect of female sex compared to males does not contain zero, supporting its statistical significance. The age-centered age interval also excludes zero, consistent with the original finding.

In summary, using 100 replicates, these bootstrap results generally corroborate the direction, magnitude, and statistical significance of the fixed effects estimated in the final refined excess cost model. This provides increased confidence in the stability of the findings regarding the persistent excess healthcare costs associated with Lyme disease diagnosis.

8 Comparison with other diagnoses

To better understand the significance of the healthcare cost patterns associated with Lyme disease, it is informative to compare its cost trajectory relative to diagnosis with those of other distinct medical conditions. This comparative analysis helps contextualize the economic impact of Lyme disease, highlighting similarities or differences in the cost patterns following diagnosis compared to conditions ranging from acute infections to major chronic diseases.

For this, the same models for other diseases were created as described in Chapter 7.2. Figure 7 illustrates the average impact of the disease on annual healthcare costs for several conditions in the year of diagnosis and the three subsequent years, based on the models output.

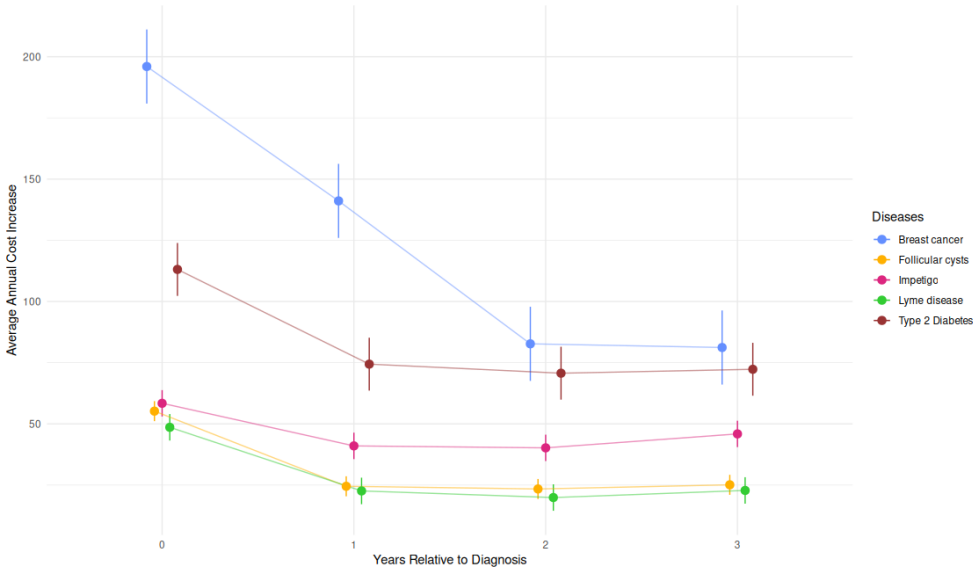

Figure 7: Impacts of different diseases on the average yearly healthcare costs (in euros)

Figure 7 reveals markedly different cost profiles across the selected diseases. Conditions such as breast cancer and type 2 diabetes exhibit substantially higher costs, particularly around the diagnosis year. Breast cancer shows exceptionally high initial costs that decrease significantly but remain elevated, likely reflecting intensive

initial treatment followed by ongoing care. Type 2 diabetes shows high initial costs followed by a plateau at a persistently high level, consistent with its chronic nature requiring continuous management.

In contrast, conditions typically considered acute, such as impetigo (a bacterial skin infection) and follicular cysts of skin, display much lower overall costs. Their cost trajectories are characterized by a peak near the diagnosis year followed by a rapid decline to a low baseline in subsequent years. This pattern suggests that the use of healthcare resources is concentrated around the acute phase of diagnosis and treatment, with minimal lingering costs. For more detailed tables on estimates, see Appendix 2.

While the peak cost in the diagnosis year is considerably lower than for breast cancer or type 2 diabetes, it is comparable to the costs observed for follicular cysts and impetigo in this dataset. However, impetigo remains generally with higher costs over the longer period. This pattern contrasts with the rapid decline to a lower baseline observed for Lyme disease and follicular cysts.

This persistence of moderately elevated costs for Lyme disease in the years following diagnosis, preventing a return to a low baseline similar to impetigo or follicular cysts, warrants attention. It suggests an economic burden that, on average, extends beyond the acute treatment phase. This pattern could potentially reflect costs associated with several factors known to be relevant in Lyme disease, such as follow-up consultations, management of incomplete symptom resolution, diagnosis and treatment of later-stage manifestations (for example Lyme arthritis), or healthcare interactions related to persistent non-specific symptoms often discussed in the context of Post-Treatment Lyme Disease Syndrome. While this analysis visualizes average trends and cannot pinpoint specific causes for individual costs, the overall trajectory strongly indicates that Lyme disease alongside with other diseases can cause elevated healthcare costs for prolonged periods.

It is essential to interpret these findings considering they represent average cost

trends. Furthermore, the costs depicted reflect direct healthcare expenditures captured in the data source and do not encompass indirect costs, such as productivity losses, or potential prescriptions that need to be bought out. Nevertheless, these results underscore the unique economic profile of Lyme disease, less costly than major chronic illnesses or cancers, but potentially imposing a sustained burden on diagnosed people.

Conclusion

This thesis aimed to measure how healthcare costs change over time when someone is diagnosed with Lyme disease, using data from the Estonian Biobank and Linear Mixed-Effects models. The results showed a significant cost impact: average costs clearly peaked in the diagnosis year and remained higher than expected for at least three years afterwards, compared to before diagnosis or to people without the diagnosis. This cost pattern is similar to follicular cysts, but is also distinct from major long-term diseases known for very high ongoing costs, such as type 2 diabetes or breast cancer.

These findings highlight that the financial burden of Lyme disease often lasts longer than just the initial illness phase, which is important information for healthcare planning and resource allocation in Estonia. Methodologically, the study confirmed that LME models are a suitable and valuable tool for analyzing the kind of longitudinal cost data used here, effectively handling the correlations within individual patient records.

However, it's important to consider the study's limitations. The analysis relied on administrative health data, which is currently missing pharmaceutical costs. Other unmeasured factors related to patient health or lifestyle could also influence individual costs. Additionally, the study focused on a specific time window (-2 to +3 years relative to diagnosis), and the bootstrap validation performed to check result stability used 100 replicates, meaning it provides preliminary support.

Assigning index years randomly to non-diagnosed individuals may not perfectly replicate the calendar year distribution of Lyme diagnoses, this approach was chosen to preserve simplicity and computational efficiency within the scope of a Master's thesis. Alternative methods such as stratified sampling or matched controls by age and calendar time could reduce potential residual confounding due to time trends or demographic imbalance. However, these would introduce considerable added complexity. Moreover, demographic variables such as age and sex were in-

cluded as covariates in the modeling to partially mitigate these concerns.

Future research could build on these findings by exploring cost trends over a longer follow-up period, if data permits. Including more detailed clinical information, such as disease severity or specific treatments received, could help explain the observed differences between patients. Breaking down the total costs into specific categories could also offer deeper insights into how healthcare resources are used following a Lyme disease diagnosis.

In summary, this thesis provides solid quantitative evidence that the diagnosis of Lyme disease is associated with a significant and lasting increase in direct healthcare costs within the Estonian context. By characterizing the specific timing and persistence of this impact, the research offers useful information for public health planning and contributes to a better understanding of the economic consequences of this common disease.

Acknowledgements

The activities of the Estonian Biobank are strictly regulated by the Estonian Human Genes Research Act. The research carried out for this thesis, involving individual-level data analysis, was carried out under ethical approval 1.1-12/624 granted by the Estonian Committee for Bioethics and Human Research. Data access and use were further governed by data release application 6-7/GI/33520 from the Estonian Biobank.

We want to acknowledge the Estonian Biobank Research Team, who was responsible for data collection, genotyping, quality control, and imputation and consisted of Andres Metspalu, Mait Metspalu, Lili Milani, Reedik Mägi, Mari Nelis, and Georgi Hudjashov. We also kindly thank Andres Võrk for providing us with healthcare cost information.

The analyses were carried out in the High Performance Computing Center, University of Tartu.

Literature

Borchers, A.T., Keen, C. L., Huntley, A. C., and Gershwin, M. E. (2015). "Lyme disease: A rigorous review of diagnostic criteria and treatment". In: *Journal of Autoimmunity* 57, pp. 82–115. ISSN: 0896-8411. DOI: https://doi.org/10.1016/j.jaut.2014.09.004.

Bratton, R. L., Whiteside, J. W., Hovan, M. J., Engle, R. L, and Edwards, F. D. (2008). "Diagnosis and treatment of Lyme disease". In: *Mayo Clinic Proceedings*. Vol. 83. 5. Elsevier, pp. 566–571.

Burgdorfer, W., Hayes, S. F., and Corwin, D. (1989). "Pathophysiology of the Lyme disease spirochete, Borrelia burgdorferi, in ixodid ticks". In: *Reviews of infectious diseases* 11, S1442–S1450.

Bush, L. M. and Vazquez-Pertejo, M. T. (2018). "Tick borne illness—Lyme disease". In: *Disease-a-Month* 64.5. SI: SELECTED EPIDEMICS and EMERGING PATHOGENS – VECTOR BORNE ILLNESSES, pp. 195–212. ISSN: 0011-5029. DOI: https://doi.org/10.1016/j.disamonth.2018.01.007.

Centers for Disease Control and Prevention (CDC) (May 15, 2024). *Lyme Disease Rashes*. URL: https://www.cdc.gov/lyme/signs-symptoms/lyme-disease-rashes.html (visited on May 15, 2025).

Efron, B. (1979). "Bootstrap Methods: Another Look at the Jackknife". In: *The Annals of Statistics* 7.1, pp. 1–26. DOI: 10.1214/aos/1176344552.

Embers, M. E. and Narasimhan, S. (2013). "Vaccination against Lyme disease: past, present, and future". In: *Frontiers in cellular and infection microbiology* 3, p. 6.

European Centre for Disease Prevention and Control (ECDC) (2016). *Factsheet about Lyme borreliosis*. URL: https://www.ecdc.europa.eu/en/borreliosis/facts/factsheet (visited on May 19, 2025).

Fisher, R. A. (1918). "XV.—The Correlation between Relatives on the Supposition of Mendelian Inheritance." In: *Transactions of the Royal Society of Edinburgh* 52.2, pp. 399–433. DOI: 10.1017/S0080456800012163.

Harville, D. A. (1977). "Maximum likelihood approaches to variance component estimation and to related problems". In: *Journal of the American Statistical Association* 72.358, pp. 320–338. DOI: 10.2307/2286796.

Hatchette, T. F., Davis, I., and Johnston, B. L. (2014). "Lyme disease: clinical diagnosis and treatment". In: *Canada Communicable Disease Report* 40.11, p. 194.

Krupkaa, M., Raskaa, M., Belakovaa, J., Horynovaa, M., Novotnyb, R., and Weigla, E. (2007). "Biological aspects of Lyme disease spirochetes: unique bacteria of the Borrelia burgdorferi species group." In: *Biomedical Papers of the Medical Faculty of Palacky University in Olomouc* 151.2.

Kullberg, B. J., Vrijmoeth, H. D., Schoor, F. van de, and Hovius, J. W. (2020). "Lyme borreliosis: diagnosis and management". In: *BMJ* 369.

Laird, N. M. and Ware, J. H. (1982). "Random-effects models for longitudinal data". In: *Biometrics* 38.4, pp. 963–974. DOI: 10.2307/2529876.

Leeflang, M. M. G., Ang, C. W., Berkhout, J., Bijlmer, H. A., Van Bortel, W., Brandenburg, A. H., Van Burgel, N. D., Van Dam, A. P., Dessau, R. B., Fingerle, V., et al. (2016). "The diagnostic accuracy of serological tests for Lyme borreliosis in Europe: a systematic review and meta-analysis". In: *BMC infectious diseases* 16, pp. 1–17. DOI: https://doi.org/10.1186/s12879-016-1468-4.

Maksimyan, S., Syed, M. S., and Soti, V. (2021). "Post-treatment Lyme disease syndrome: need for diagnosis and treatment". In: *Cureus* 13.10. DOI: 10.7759/cureus.18703.

Marques, A. R. (2015). "Laboratory diagnosis of Lyme disease-advances and challenges". In: *Infectious disease clinics of North America* 29.2, p. 295.

Nguyen, C. T., Cifu, A. S., and Pitrak, D. (Feb. 2022). "Prevention and Treatment of Lyme Disease". In: *JAMA* 327.8, pp. 772–773. ISSN: 0098-7484. DOI: 10.1001/jama.2021.25302.

Parm, Ü., Niitvägi, E., Beljaev, K., Aro, T., Aotäht, E., Raska, K., Epštein, J., Oona, M., and Lutsar, I. (2015). "Puukborrelioos Saaremaal". In: *Eesti Arst*. DOI: 10.15157/ea.v0i0.12011.

Pinheiro, J. C. and Bates, D. M. (2000). "Linear mixed-effects models: Basic concepts and examples". In: *Mixed-Effects Models in S and S-PLUS*. Springer, pp. 3–56. DOI: https://doi.org/10.1007/b98882.

Prükk, T., Maimets, M., and Lutsar, I. (2012). "Lyme'i tõve nüüdisaegne diagnostika ja ravi". In: *Eesti arst*.

Rebman, A. W., Crowder, L. A., Kirkpatrick, A., and Aucott, J. N. (2015). "Characteristics of seroconversion and implications for diagnosis of post-treatment Lyme disease syndrome: acute and convalescent serology among a prospective cohort of early Lyme disease patients". In: *Clinical rheumatology* 34, pp. 585–589.

Robinson, M. L., Kobayashi, T., Higgins, Y., Calkins, H., and Melia, M. T. (2015). "Lyme carditis". In: *Infectious disease clinics of North America* 29.2, pp. 255–268.

Roome, A., Spathis, R., Hill, L., Darcy, J. M., and Garruto, R. M. (2018). "Lyme disease transmission risk: seasonal variation in the built environment". In: *Healthcare*. Vol. 6. 3. MDPI, p. 84.

Russell, A. L. R., Dryden, M. S., Pinto, A. A., and Lovett, J. K. (2018). "Lyme disease: diagnosis and management". In: *Practical neurology* 18.6, pp. 455–464.

Sala, V., De Faveri, E., et al. (2016). "Epidemiology of Lyme disease in domestic and wild animals". In: *The Open Dermatology Journal* 10.Suppl. 1: M3, pp. 15–26.

Schoen, R. T. (2020). "Challenges in the diagnosis and treatment of Lyme disease". In: *Current rheumatology reports* 22, pp. 1–11.

Shapiro, E. D. (2014). "Lyme Disease". In: *New England Journal of Medicine* 370.18, pp. 1724–1731. ISSN: 0028-4793. DOI: 10.1056/NEJMcp1314325.

Sprong, H., Azagi, T., Hoornstra, D., Nijhof, A. M., Knorr, S., Baarsma, M. E., and Hovius, J. W. (2018). "Control of Lyme borreliosis and other Ixodes ricinus-borne diseases". In: *Parasites & vectors* 11, pp. 1–16.

Stanek, G., Wormser, G. P., Gray, J., and Strle, F. (2012). "Lyme borreliosis". In: *Lancet (London, England)* 379.9814, pp. 461–473. DOI: 10.1016/S0140-6736(11)60103-7.

Sykes, J. E. and Allen, J. L. (2011). "Canine and Human Lyme Borreliosis". In: *Veterinary Clinics of North America: Small Animal Practice* 24.3. Provides broad context on LB prevalence., pp. 1189–1215. ISSN: 0891-5520. DOI: 10.1016/j.cvfa.2011.04.006.

Terviseamet (2025). *Infectious diseases transmitted by ticks*. URL: https://www.terviseamet.ee/nakkushaigused/puugihaigused#puugihaigused-ja-enn (visited on Apr. 5, 2025).

Voyiatzaki, C., Papailia, S. I., Venetikou, M. S., Pouris, J., Tsoumani, M. E., and Papageorgiou, E. G. (2022). "Climate changes exacerbate the spread of Ixodes ricinus and the occurrence of Lyme borreliosis and tick-borne encephalitis in Europe—how climate models are used as a risk assessment approach for tick-borne diseases". In: *International journal of environmental research and public health* 19.11, p. 6516.

Wong, K. H., Shapiro, E. D., and Soffer, G. K. (2022). "A review of post-treatment Lyme disease syndrome and chronic Lyme disease for the practicing immunologist". In: *Clinical reviews in allergy & immunology* 62.1, pp. 264–271.

Wright, W. F., Riedel, D. J., Talwani, R., and Gilliam, B. L. (2012). "Diagnosis and management of Lyme disease". In: *American family physician* 85.11, pp. 1086–1093.

Appendix 1. Non-diagnosed group model estimates

Table 7: Non-diagnosed group model estimates (Year 0 as Ref.)

Predictor	**Estimate**	**Std. Error**	**p-value**
Intercept	55.979	0.732	<0.001
Baseline costs	0.348	0.002	<0.001
Year 1	10.476	0.629	<0.001
Year 2	20.561	0.629	<0.001
Year 3	34.244	0.629	<0.001
Sex (Female vs Male Ref.)	21.663	0.740	<0.001
Age (centered at 40)	1.667	0.021	<0.001

Appendix 2. Additional diseases models impact estimates

Table 8: Fixed effects estimates: adjusted excess cost model for Breast Cancer

Predictor	**Estimate**	**Std. Error**	**p-value**
Year 0	195.965	7.722	<0.001
Year 1	141.071	7.722	<0.001
Year 2	82.723	7.722	<0.001
Year 3	81.244	7.722	<0.001

Table 9: Fixed effects estimates: adjusted excess cost model for Type 2 Diabetes

Predictor	**Estimate**	**Std. Error**	**p-value**
Year 0	113.131	5.512	<0.001
Year 1	74.353	5.512	<0.001
Year 2	70.683	5.512	<0.001
Year 3	72.325	5.512	<0.001
Sex (Female vs Male Ref.)	-19.623	4.888	<0.001
Age (centered at 40)	0.549	0.179	0.002

Table 10: Fixed effects estimates: adjusted excess cost model for Follicular Cysts

Predictor	**Estimate**	**Std. Error**	**p-value**
Year 0	55.260	2.082	<0.001
Year 1	24.504	2.082	<0.001
Year 2	23.374	2.082	<0.001
Year 3	25.131	2.082	<0.001
Age (centered at 40)	0.695	0.089	<0.001

Table 11: Fixed effects estimates: adjusted excess cost model for Impetigo

Predictor	**Estimate**	**Std. Error**	**p-value**
Year 0	58.405	2.751	<0.001
Year 1	41.031	2.751	<0.001
Year 2	40.220	2.751	<0.001
Year 3	45.865	2.751	<0.001
Age (centered at 40)	1.136	0.101	<0.001

Non-exclusive licence to reproduce the thesis and make the thesis public

I, Sten Raak,

1. grant the University of Tartu a free permit (non-exclusive licence) to reproduce, for the purpose of preservation, including for adding to the digital archives of the University of Tartu until the expiry of the term of copyright, my thesis "LYME DISEASE: MODELING AND ANALYZING LONG-TERM COSTS RELATED TO THE INFECTION BASED ON ESTONIAN BIOBANK DATA", supervised by Erik Abner and Krista Fischer;

2. grant the University of Tartu a permit to make the thesis specified in point 1 available to the public via the web environment of the University of Tartu, including via the digital archives, under the Creative Commons licence CC BY NC ND 4.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work until the expiry of the term of copyright;

3. am aware of the fact that the author retains the rights specified in points 1 and 2;

4. confirm that granting the non-exclusive licence does not infringe other persons' intellectual property rights or rights arising from the personal data protection legislation.

Sten Raak
21.05.2025